KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents

Tom Magerman, Bart Van Looy, Bart Baesens and Koenraad Debackere

DEPARTMENT OF MANAGERIAL ECONOMICS, STRATEGY AND INNOVATION (MSI)

OR 1114

Assessment of Latent Semantic Analysis (LSA) text mining algorithms for large scale mapping of patent and scientific publication documents.

Tom Magerman

Bart Van Looy

Bart Baesens

Koenraad Debackere

October 2011

Abstract

In this study we conduct a thorough assessment of the LSA text mining method and its options (preprocessing, weighting, …) to grasp similarities between patent documents and scientific publications to develop a new method to detect direct science-technology linkages - as this is instrumental for research on topics in innovation management, e.g. anti-commons issues. We want to assess effectiveness (in terms of precision and recall) and derive best practices on weighting and dimensionality reduction for application on patent data. We use LSA to derive similarity from a large set of patent and scientific publication documents (88,248 patent documents and 948,432 scientific publications) based on 40 similarity measurement variants (four weighting schemas are combined with ten levels of dimensionality reduction and the cosine metric). A thorough validation is set up to compare the performance of those measure variants (expert validation of 300 combinations plus a control set of 30,000 patents). We do not find evidence for the claims of LSA to be superior to plain cosine measures or simple common term or co-occurrence based measures in our data; dimensionality reduction only seems to approach cosine measures applied on the full vector space. We propose the combination of two measures based on the number of common terms (weighted by the minimum of the number of terms of both documents and weighted by the maximum of the number of terms of both documents respectively) as a more robust method to detect similarity between patents and publications.

1 Introduction

The link between science and technology, and more particular, the interaction of science and technology and resulting impact on scientific and technological development and welfare, is subject to many studies on innovation and economic development. Researchers in the domain of innovation (see e.g. Freeman, 1987 and 1994; Lundvall, 1992; Nelson, 1993; Nelson & Rosenberg, 1993; Mansfield, 1995; Mansfield & Lee, 1996; Mowery & Nelson, 1999; Dosi, 2000) stress the role of science and the importance of interaction between a variety of institutional actors underlying the innovative capacity and consequent economic performance of an economical system, and the importance of the institutional framework (see e.g. the Triple Helix model: Leydesdorff & Etzkowitz, 1996 and 1998; Etzkowitz & Leydesdorff, 1997).

Interaction and exchange between academia and industry can result in positive aspects, both for the business partner (e.g. Zucker & Darby, 2001; Hall, Link & Scott, 2001; Faems, Van Looy & Debackere, 2005) and for the academic sector (e.g. realization of complementarities between applied and basic research – Azoulay, Ding & Stuart, 2009; generation of new research ideas – Rosenberg, 1998; attracting additional resources for (basic) research - Agrawal & Henderson, 2002). Additional benefits – when introducing intellectual property in scientific activities - can be found in the facilitation of the creation of a market for ideas and the ability of society to realize the commercial and social benefits of a given discovery (Kitch, 1977; Merges & Nelson, 1990; Gans & Stern, 2000; Arora, Fosfuri & Gambardella, 2004; Hellman, 2007; Murray & Stern, 2007).

At the same time some concerns arise due to the increasing commercialization of scientific activities undertaken by universities. Too much emphasis on (market) exploitation might negatively impact the quantity and quality of scientific research and change research orientation because of changing incentives (skewing problem: research topic decisions follow market demand and money). But also indirect effects get attention: shift of career choices of promising young graduate students and post-doctorals away from academia; increasing secrecy or delay of publication (demanded by industrial partners); and presence of an anti-commons effect (to many owners

blocking the use of inventions). These concerns get particular attention because they might slow down the rate of innovation and long-term scientific and technological advancement might be traded in for short term benefits.

While most empirical evidence – at the level of individual scientists – reports a positive relationship between patenting activities and publication outcomes (quantity as well as quality), expansion of IPR might result in 'privatizing' the scientific commons and potentially limiting scientific progress (Argyres & Liebeskind, 1998; David, 2000; Krimsky, 2004). This fear is nicely expressed by the metaphor of the 'Tragedy of the anti-commons', introduced by Heller (Heller, 1998) as opposed to the 'Tragedy of the commons' of Hardin (Hardin, 1968). Heller states that the presence of too many owners with blocking power can lead to the underutilization of scarce resources, or, translated to the world of IPR, more intellectual property rights may lead paradoxically to fewer useful products instead of being an incentive to invent and disclose (too many owners hold rights in previous discoveries creating obstacles for future research).

Although anecdotal evidence exists of problematic impact of IPR on scientific findings (e.g. the 'OncoMouse' or 'Havard mouse' of Leder and Stewart; and patents on human genes associated with breast and ovarian cancer owned by Myriad Genetics), large scale evidence of the presence of an anti-commons effect in biotechnology patenting is rare and the magnitude of the phenomenon and the real thread of patent thickets to block access to knowledge and technology is unclear. One notable exception is the study of Murray & Stern (2007) suggesting a modest anti-commons effect based on a decline in citation rate – after granting of the patent - by 10 to 20% for a set of 169 patent-publication pairs published in Nature Biotech between 1997 and 1999, although these authors also clearly point to the interpretation limits inherent to their study.

A major challenge for the study of the presence of an anti-commons effect, and in studies on science and technology interactions in general, is the identification of science-related patents in general and the identification of scientific results protected by intellectual property rights (IPR) in particular to understand the magnitude and characteristics of the phenomenon. For broader or high-level studies at the level of countries, sectors or technologies, the matching of non-patent references with

databases with bibliographic data or scientific publications might yield valuable insights. For more low-level studies or direct science-technology interactions, current approaches involve the use of the number of non-patent references on patent documents, or the matching of patent inventor names and patentees with publication authors and affiliations. The former approach based on the number of non-patent references is easy to conduct on a large scale but suffers from the vagueness of the value of a non-patent reference as an indicator of science-intensity or science-relatedness (see e.g. Callaert, Van Looy et al., 2006). The latter approach based on patentee and inventor name matching is more robust, but is not easy to implement on a large scale: patentee name matching requires name cleaning and addressing problems of name changes, name variants and organization entity resolution (from research groups to faculties/departments and institutions/universities); inventor name matching requires dealing with homonyms and first names and middle names and initials.

A promising new approach involves text mining to directly match text documents based on their contents to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address, as this might allow (semi)-automated compilation of large datasets based on content similarity. In general this could be instrumental for, amongst others, domain studies, trend detection/emerging field detection and science-technology linkage and thus contribute to technology and innovation research. At this moment, we are particularly interested in this text mining approach to identify patents related to scientific publications based on their shared contents and especially to check for documents with identical contents to identify scientific publications protected by patents, allowing to compile large datasets to check for the presence of an anti-commons effect.

However, applicability of off-the-shelve text mining solutions is not straightforward at this stage. Multiple methods are available but existing experience for patent data is limited and more research is needed concerning effectiveness and best practices (methods, pre-processing, source data, indexing options, number of concepts to be retained, …). We focus on one method, Latent Semantic Analysis (LSA) - a statistical

method to match text documents that involves Singular Value Decomposition (SVD) for dimensionality reduction - that has proven to be effective in deriving topical relations in large sets of natural language text documents in some context (see Landauer, McNamara et al., 2007, for a detailed description of the method and an overview of applications). Our earlier experience with LSA and patent data proved that results are promising but also that generally accepted LSA options do not always yield the best results with patent data (Magerman, Van Looy & Song, 2010).

In this study we conduct a thorough assessment of the LSA text mining method and its options (preprocessing, weighting, …) to grasp similarities between patent documents and scientific publications. We want to assess effectiveness (in terms of accuracy/precision and exhaustiveness/recall) and derive best practices on weighting and dimensionality reduction for application on patent data, given the technical and juridical nature and hence different linguistic context of patent and scientific publication documents. Our primary goal is to set up a method to identify scientific publications that are protected by patents (so called 'patent-publication' or 'patent-paper' pairs, i.e. scientific publications from which the contents – methodology, findings, discovery/invention – is part of a patent publication). We use LSA to derive similarity from a large set of patent and scientific publication documents (88,248 patent documents and 948,432 scientific publications) based on 40 similarity measurement variants; four weighting schemas – no term weighting; binary weighting; inverted document frequency; and term frequency x inverted document frequency – are combined with ten levels of dimensionality reduction – no SVD reduction; 1,000; 500; 300; 200; 100; 50; 25; 10; 5 dimensions – and the cosine metric. In addition we also include three similarity measure variants into the comparison based on the number of common terms weighted by the total number of terms of the documents. A thorough validation is set up to compare the performance of those measure variants: the degree of similarity of 300 patent-publication combinations is rated by experts to compare with the outcomes of the text mining measures and about 30,000 patents from control sets are used to check the robustness of the expert validation results.

We do not find evidence for the claims of LSA to be superior to simple common term or co-occurrence based measures for content similarity in our data. At the contrary, a cosines metric applied to the full vector space or term-by-document matrix, without any dimensionality reduction, clearly outperforms LSA based measures, as well as simple measures based on the number of common terms. The term weighting method used also has considerable impact; binary and IDF weighting yields better results compared to TF-IDF weighting and no weighting at all, a remarkable observation as TF-IDF in combination with SVD retaining 300-500 dimensions is a commonly used method. We propose the combination of two measures based on the number of common terms (weighted by the minimum of the number of terms of both documents and weighted by the maximum of the number of terms of both documents respectively) as a more robust method to detect similarity.

We continue this publication with a brief introduction to text mining and the LSA method. Next we describe our patent and publication data and our method to match patents and publications. In section 4 we dig into our derivation of content similarity and describe our 43 measure variants to compare. Section 5 presents first aggregative results, followed by a first expert validation in section 6, control set validation in section 7 and final expert validation and results in section 8. In section 9 we try to shed a light on the poor performance of LSA-based measures. We round up with conclusions and directions for further research in section 10.

2 Introduction to text mining, potential applications in the field of innovation studies, and the Latent Semantic Analysis (LSA) method.

2.1 Text mining

Text mining refers to the automated extraction of knowledge and information from text by means of revealing relationships and patterns present, but not obvious, in a document collection. Text mining covers a broad field of tasks including text categorization, text clustering, information extraction, sentiment analysis, document summarization, named entity recognition and question answering and is an

interdisciplinary field based on artificial intelligence, natural language processing, computational linguistics, information retrieval, data mining, machine learning and statistics.[1]

Technological advances and large scale availability of computing power attracted a lot of interest for text mining in recent years, together with the observation that the vast majority of (electronically available) information is stored as (unstructured) text and not in structured databases. Hence database technologies and knowledge discovery in structured databases (data mining) alone will fall short to disclose knowledge and information from available resources. Text mining techniques can help to reveal knowledge and information from large text collections, disclosing data not available in structured databases. Given the overwhelming amount of unstructured data recorded as texts, text mining will become increasingly valuable for research.

It is important to stress that these text mining techniques go beyond information retrieval. Information retrieval helps in finding information based on a user request, and is it obvious that text mining techniques can help in improving this. As such, currently, information retrieval is probably the biggest area of text mining application and related techniques are widely used. Information retrieval in itself does however not discover new knowledge or insights, it just reveals what is already known to somebody (and also the user issuing the search request has to know what he is looking for)[2].

Text mining does go one step further and is about knowledge discovery, revealing new things that were not obvious to discover by humans. Nice illustrations are some cases of a literature-based approach to scientific discovery by Swanson: fish oil and Raynaud's syndrome; migraine and magnesium; and somatomedin C and arginine (Swanson, 1986, 1988 and 1990). The second case e.g. describes the discovery of the relationship between 'migraine' and 'spreading depression' on the one hand, and 'magnesium' and 'preventing depression' on the other hand after a thorough search into medical literature on migraine, suggesting magnesium deficiency as a factor in migraine. Prior to

[1] For more information on the application of text mining and its relation to other fields and techniques, see e.g. Hearst, 1999, or Fan, Wallace et al., 2006. For an overview of techniques, see Vidhya & Aghila, 2010.

[2] For an elaboration on this issue, see Hearst, 1999.

this remarkable discovery – Swanson is an information scientist, not a physician - there was no indication of this relationship whatsoever; his results triggered additional clinical research, confirming his suggestion[3]. These case studies can be regarded as pioneering cases of text mining – when text mining as such did not even exist – and were at the basis of formalized study to literature-based discovery – so called Swanson Linking (Swanson & Smalheiser, 1997).

2.2 History

Quantitative linguistics dates back to at least the middle of the 19th century (see Grzybek & Kelih, 2004). However, the classical theoretical work by Zipf (1949) is considered pioneering in quantitative linguistic (or text) analysis. Since the 1970s, a remarkable increase in activity has been witnessed in this aspect of information science. As for its application to scientific literature, Wyllys's study (1975) is among the first. Co-word analysis, one of the most frequent techniques, was founded on the idea that the co-occurrence of words describes the contents of documents and was developed for purposes of evaluating research (Callon, Courtial et al., 1983). The extension of co-word analysis to the full texts of large sets of publications was possible as soon as large textual databases became available in electronic form; also the increasing availability of computational power allowed further emergence of text mining approaches. Manning & Schütze (2000) provide a comprehensive introduction to the statistical analysis of natural language; Berry (2003) provide a survey of text mining research; Leopold, May & Paaß (2004) give an overview of data and text mining fundamentals for science and technology research; and Porter & Newman (2004) introduced the term 'tech mining' to text mining of collections of patents on a specific topic. Other practical applications in the field of bibliometrics and technometrics are presented by Courtial (1994), Noyons, van Raan et al. (1994), Bassecoulard & Zitt (2004), Leydesdorff (2004), Glenisson, Glänzel et al. (2005) and Janssens, Leta et al. (2006).

[3] Ramadan, Halvorson et al., 1989

2.3 Application in innovation studies

As described in the previous section, application of text mining techniques in innovation studies is not new and can provide the necessary input for a complex research setup that would be impossible or at least difficult (i.e. time consuming because of involved manual data treatment) to conduct without text mining techniques.

As mentioned earlier, a first and more traditional application of text mining is in the field of information retrieval (conducting patent or publication searches on bibliographic databases). But text mining techniques also allow for a new range of applications:

- Domain studies: starting from a set of 'seed patents' that are representative for a technological domain, concepts and topics can be extracted and used to match with concepts and topics of other patents to identify related patents and delineate technological domains by a set of patents;
- Trend detection / emerging field detection: Concepts and topics extracted from a set of patent documents can be clustered over time to identify new domains or to follow the evolution of a domain;
- Science-technology linkage: concepts and topics extracted of a set of patent documents can be compared to concepts and topics extracted from a set of scientific publications to reveal similarity between patents and publications.

2.4 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) was developed late 1980s at BellCore/Bell Laboratories by Landauer and his team of Cognitive Science Research (Landauer & Dumais, 1997):

"Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the meaning of words. Meaning is estimated using statistical computations applied to a large corpus of text. The corpus embodies a set of mutual constraints that largely

determine the semantic similarity of words and sets of words. These constraints can be solved using linear algebra methods, in particular, Singular Value Decomposition."[4]

LSA is a mathematical and statistical approach, claiming that semantic information can be derived from a word-document co-occurrence matrix and words and documents can be represented as points in a (high-dimensional) Euclidean space. Dimensionality reduction is an essential part of this derivation.

LSA is based on the Vector Space Model (VSM), an algebraic representation of text documents commonly used in information retrieval. This 'bag-of-words' approach can be seen as a simple yet powerful representation (Salton, 1968; Salton, Wong & Yang, 1975; Salton & McGill, 1983). The vector space of a collection of texts is constructed by representing each document as a vector containing the frequencies of the words or terms the document is composed of as elements. Altogether, these document vectors add up to a term-by-document matrix representing the full text collection. Relatedness of documents can be derived from those vectors, e.g. by calculating the angle between document vectors by means of a cosine measure.

However, this numerical representation of text data does not solve typical issues of working with language. On the one hand there are morphological problems for the proper identification of terms and the fact that not all terms in a text are of equal importance. This can be solved by feature selection techniques (stemming, stop word removal, collocations, synonym lists, domain vocabulary, part-of-speech taggers, chi-square tests and information gain) and weighting schemes (TF-IDF, Log-Entropy). On the other hand, there are or more fundamental issues with homonymy/polysemy and synonymy. These issues require specific methods to (try to) understand the meaning of words, and that is what LSA claims to do:

"It was thus a major surprise to discover that a conceptually simple algorithm applied to bodies of ordinary text could learn to match literate humans on tasks that if done by

[4] Landauer, McNamara et al. (Eds.), 2007, Handbook of Latent Semantic Analysis, Preface x.

people would be assumed to imply understanding of the meaning of words and passages. The model that first accomplishes this feat was LSA."[5]

LSA rests on the single conceptually simple constraint that the representation of any meaningful passage of text must be composed as a function of the representation of the words it contains. Thus, LSA models a passage as a simple linear equation, and a large corpus of text as a large set of simultaneous equations. The solution is in the form of a set of vectors, one for each word and passage, in a semantic space and is solved by Singular Value Decomposition (SVD).

Optimal dimension reduction is a common workhouse in analysis of complex problems in many fields of science and engineering. Over 99.9% of the cells in the word-by-paragraph or term-by-document matrix representing the documents in the vector space turn out to be empty. This makes the comparison of word or paragraph meanings quite chancy. However, after dimension reduction and reconstruction, every cell will be filled with an estimate that yields a similarity between any paragraph and any other and between any word and any other. This dimension reduction is crucial and is what accounts for LSA's advantage over most current methods of information retrieval, which rely on matching literal words. It is also what accounts for its ability to measure the similarity of two essays that use totally different words, and for all of the other properties of LSA that defy the intuition that learning language from language is impossible.

LSA builds upon semantic similarity and hence uses proximity models such as clustering, factor analysis and multidimensional scaling (see Carroll & Arabie, 1980, for a survey). Discovering latent proximity structure has previously been explored for automatic document indexing and retrieval, using term and document clustering (Sparck Jones, 1971; Salton, 1968; Jardin & van Rijsbergen, 1971) and factor analysis (Atherton & Borko, 1965; Borko & Bernick, 1963; Ossorio, 1966); LSA builds further on these factor analysis techniques and constructs a concept-by-document matrix using a low-rank

[5] Landauer, McNamara et al. (Eds.), 2007, Handbook of Latent Semantic Analysis, page 5.

approximation of the term-by-document matrix, combining dimensions or terms into 'concepts'.

Singular Value Decomposition (SVD) is used as a rank lowering method to truncate the original vector space to reveal the underlying or 'latent' semantic structure in the pattern of word usage to define documents in a collection. This truncation allows dealing with typical language issues like synonymy as different words expressing the same idea are supposed to be close to each other in the reduced *k*-dimensional vector space. SVD will decompose the original term-by-document matrix into orthogonal factors that represent both terms and documents:

$$A = U \,.\, \Sigma \,.\, V^T$$

with A the original term-by-document matrix, Σ a diagonal matrix with the square roots of singular values of $A.A^T$ and $A^T.A$ ($\sigma_1^2 > \sigma_2^2 > \ldots > \sigma_n^2$), and U and V containing left and right singular vectors.

Instead of working with the original vector space represented by the original term-by-document matrix A, one can work with the reduced vector space of lower rank, ignoring all but the first *k* singular values in Σ and all but the first *k* columns of U and V:

$$A = A^{m\times n} \cong A_k^{m\times n} = U^{m\times k}.\Sigma^{k\times k}.V^{k\times n}$$

This dimension reduction to *k* dimensions provided by SVD is the closest rank-*k* approximation available and allows eliminating noise and capturing the underlying latent structure. The *k* dimensions in the new space are no longer (stemmed) words or terms, but linear combinations of such linguistic terms, and the basic unit of analysis becomes not just a mere word but a word-and-its-context, a concept (hence the denomination of 'concept space').

Mind that the dimension reduction is not about computational simplification[6] but a fundamental aspect of the method to deal with language issues and reduce noise (terms

[6] At the contrary, SVD will convert the original sparse matrix into a full matrix. Even for low values of *k* – the number of retained dimension or concepts – this will result in a new document-by-term matrix of

in documents that do not contribute to the meaning of the document or parts of the document). As such, the choice of *k* is not arbitrary but needs to be chosen carefully to truly represent the underlying latent structure of the data.

The choice of the number of concepts to be retained is not straightforward. Current literature suggests to take 100 to 300 concepts for large datasets (Berry, Drmac & Jessup, 1999; Jessup & Martin, 2001; Lizza & Sartoretto, 2001). For some applications it might be better to use a subset of the first 100 or 300 dimensions (Landauer & Dumais, 1997).

2.5 Practical indexing and additional pre-processing steps

Indexing in practice

The encoding of the documents into vectors is called indexing. During indexing, a global vocabulary is built up, assigning a unique identifier to each word encountered in the entire document collection. With this global vocabulary, a vector is constructed for each document with as many elements as the total number of words in the global vocabulary, generating the vector space. For words appearing in the document at hand, the value of the respective vector elements of the document vector is equal to the number of occurrences of that word in the document at hand. For words not appearing in the document, the respective vector elements obtain a zero value. Thus, each document is represented by a vector representing raw frequencies of occurrences in a high-dimensional vector space of terms. As each document uses only a small subset of words to describe its content, the resulting matrix is extremely sparse (containing mostly zeros)[7].

To improve the indexing process and achieve better grasp of the context of the documents, subsequent additional pre-processing actions are commonly used:

lower rank but occupying far more memory and in general taking more computational resources to process.

[7] About 99.99% zero values.

Stop word removal

All common words that do not contribute to the distinctive meaning and context of documents can be removed before indexing (e.g. “a”, “the”). Commonly used word lists are available containing a large set of so-called ’stop’ words (e.g. the SMART list of Buckley and Salton, Cornell University).

Stemming

Instead of indexing words as they appear in the documents, linguistic stems can be used for indexing. The basic idea is to reduce the number of words by introducing a common denominator, called a stem, for words that share a common meaning (e.g. ‘produc’ for “product”, “production”, “producing”, etc.). A well-known example is the Porter stemmer (see van Rijsbergen, Robertson & Porter, 1980, and Porter, 1980). This stemmer does not perform a linguistically correct lemmatization, but takes a pragmatic approach in stripping suffixes from words to combine word variants with shared meanings.

The idea of stemming is to improve the ability to detect similarity regardless of the use of word variants (stemming reduces the number of synonyms, since multiple terms sharing the same stem are mapped onto the same concept or stem), but occasionally stemming will create new homonyms because of stemming errors[8].

Term reduction

According to Zipf’s law a large number of terms only appear in one document. Such hapaxes can be removed from the vocabulary because they are of little value in finding communality between documents.

Weighting

Representing documents based on the occurrence and co-occurrence of terms (raw frequencies) can be refined by introducing a weighting scheme to better distinguish the distinctive nature of words and terms given the specific context under study (e.g. the

[8] A more in-depth analysis of the performance and advantages and disadvantages of stemming (which are also language and corpus dependent) is outside the scope of this publication. The reader interested in this aspect is referred to Lennon, Pierce et al., 1981; Harman, 1991; Krovetz, 1995; and Porter, 2001.

word ‘computer’ does not reveal the distinctive nature of a certain contribution within a document set covering only papers on computer science). A commonly used weighting scheme is the TF-IDF weighting scheme (Salton & McGill, 1983), in which the raw term frequencies are multiplied by the inverse document frequency (IDF) for that term; this results in augmenting the impact of relatively rare terms when calculating distance measures:

$$Idf_i = \log \frac{n}{\sum_{j=1}^{n} \chi(f_{ij})},$$

with

$$\chi(t) = \begin{cases} 1 \text{ if } t > 0, \\ 0 \text{ if } t = 0 \end{cases}$$

i = term index
j = document index
f_{ij} = frequency of term i in document j
and n the number of documents.

Weighting has a similar effect as stop-word removal, since words commonly used across all documents in the document set will be down-weighted compared to medium frequency words, which carry the most significant information (Salton & Wu, 1981) – as can be expected according to Zipf’s law. On the other hand, TF-IDF weighting attributes might introduce extreme weights to words with very low frequencies. Also, TF-IDF will not grasp synonyms; hence, weights of commonly used synonyms will be over-rated, as the weights of the individual (synonym) terms will be higher than the weight of the underlying common concept. Despite these shortcomings, TF-IDF weighting is one of the most popular weighting schemes, but other weighting schemes can also be used (see Manning & Schütze, 2000, for an overview).

Additional, more advanced, pre-processing tasks can be performed to further optimize the indexing process (proper name recognition; word sense disambiguation; acronym recognition; compound term and collocation detection; feature selection using

application-specific domain vocabulary or ontology, information gain, entropy or Bayesian techniques)[9].

2.6 Similarity or distance calculation

The similarity measure typically used in information retrieval applications is the cosine similarity measure (Berry & Browne, 1999). It is an expression for the angle between vectors, formulated as an inner product of two vectors, divided by the product of their Euclidean norms.

If the vectors are normalized beforehand, this formula reduces to the simple inner product. Since, in the original vector space, all vector elements are positive (a word will appear zero times or more in a document), the results are values between 1 (for similar vectors, i.e. pointing in the same direction) and 0 (for mutually orthogonal, entirely unrelated vectors), even after application of a weighting scheme like TF-IDF. This yields distances between 0 and 1 (1 – cosine α). This no longer holds for vectors in the reduced concept space after SVD, since vector elements may become negative because of the SVD, resulting in a concept-by-document space V^{kxn} that is no longer positive semi-definite, and cosine values that might be negative, hence distances between 0 and (theoretically) 2, although values larger than 1.3 are quite rare in practice. While other similarity measures are possible (e.g. Jaccard, Dice, Euclidean distance – see Baeza-Yates & Ribeiro-Neta, 1999), the cosine measure is amongst the most commonly used when using LSA and seems superior as a similarity measure in LSA applications (Harman, 1986).

2.7 Other text mining methods

Before moving to our research setup, we wish to stress that the proposed LSA methodology is only one available method for text content based similarity deduction. Other methods e.g. do not rely on semantic representation but use semantic topic models based on generative models (e.g. probabilistic inference models like probabilistic latent semantic modelling and latent Dirichlet allocation - see e.g. Wong &

[9] A more detailed description of these topics can be found in Moens, 2006.

Yao, 1995; Hofmann, 1999; Blei, Ng & Jordan, 2003). These models do not rely on a spatial representation and do not suffer from limitations to Euclidean geometry as imposed by the assumption of LSA that documents can be represented as vectors in a vector space[10].

3 Data and methodology

3.1 Match patents and publications based on content

We primarily want to identify so called patent-publication or patent-paper pairs, i.e. scientific publications for which the contents – methodology, findings, discovery/invention – is subject of a patent application. We do this by matching patent and publication documents based on content similarity using LSA text mining algorithms. For all patents, we derive similarity scores for all publications for a set of measurements variants based on LSA. Patent-publication combinations having a high content similarity are regarded to originate from the same inventive event. The choice of the best measure to grasp meaningful relations among patent and publication documents depends on a validation exercise.

We choose biotechnology as field under study because it is an active field creating big expectations in terms of development of new economic activities and welfare creation, and because it can be labelled as an industry in which the interplay between science and technology is important. From the seventies onwards scientific findings have been playing an important role within the industry (McMillan, Narin & Deeds, 2000) resulting as well in numerous studies focusing on the role of collaboration and networking (Deeds & Hill, 1996; Baum, Calabrese & Silverman, 2000; Rothaermel & Deeds, 2004), including science-technology linkage which is particularly strong in this field (e.g. Narin & Noma, 1985; Murray, 2002; Verbeek, Callaert et al., 2002).

We compile a set of patent and publication documents related to biotechnology and calculate the content similarity between all patents and publications in the set to reveal

[10] In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of points in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007).

patent-publication combinations originating from the same inventive event. Only titles and abstracts are used as they are widely and easily available, while large sets of full-text documents are difficult or expensive to obtain.

3.2 Selection of biotechnology patents

On the patent side, we limit ourselves to the OECD definition of biotechnology to identify biotechnology patents (OECD, 2005 and 2009), defining 30 International Patent Classification subclasses/groups related to biotechnology (see Appendix 1 for the list of IPC-subclasses/groups used for the selection). We use *PATSTAT* (EPO Worldwide Patent Statistical Database) to retrieve all EPO and USPTO granted patents with application and grant year between 1991 and 2008 according to the 30 defined IPC-subclasses/groups related to biotechnology. This leads to a set of 27,241 EPO and 91,775 USPTO granted patents (*PATSTAT* edition October 2009).

As text mining techniques are applied for the further identification of patent-publication pairs, only patents with titles and a minimum abstract length of 250 characters are withheld, resulting in a final patent dataset of 7,254 EPO and 80,994 USPTO biotechnology patents (hence 88,248 patents in total).

3.3 Selection of scientific publications

On the publication side, we select biotechnology publications (articles, letters, notes, reviews)[11] from the *WOS* database (Thomson Reuters ISI Web of Science) based on the Web of Science subject classification, for the same time period 1991-2008 (volume year). 243,361 publications are revealed from subject category 'Biotechnology and Applied Microbiology'.

However, to ensure that all potentially related scientific publications are present in the dataset, we extend this 'core' publication set with publications from nine related subject categories: 'Biochemical Research Methods'; 'Biochemistry & Molecular Biology'; 'Biophysics'; 'Plant Sciences'; 'Cell Biology'; 'Developmental Biology'; 'Food

[11] Articles are by far the biggest category (90% articles compared to 1.5% letters, 2% notes and 6.5% reviews).

Sciences & Technology'; 'Genetics & Heredity' and 'Microbiology Materials'[12]. This results in more than 1.75 million additional publications for the period 1991-2008 - a considerable computational challenge for the text mining method to identify patent-publication pairs. To lower the number of publications for ease of calculations without losing too much relevant documents, we only retain those publications from this extended set that are citing or are cited by at least one publication from our core set, sizing down the extended publication set to 683,674 publications.

Finally we also add all – not necessarily biotechnology - publications from three multidisciplinary journals ('Nature', 'Science' and 'Proceedings of the National Academy of Sciences of the United States of America') resulting in 97,970 additional publications.

Again we only retain publication documents with titles and a minimum abstract length of 250 characters, resulting in a final publication set of 948,432 biotechnology related publications[13].

3.4 Selection of control sets

To check the validity of our text mining method we also compile three control sets with patent documents that are not related to biotechnology: agriculture; automotive; and materials. For each of these control sets, we randomly select 2,500 EPO and 7,500 USPTO granted patent documents from the same time period based on IPC-codes (respectively IPC class A01 for agriculture; B60 and B62 for automotive, and IPC subclass G01N, G01R and HO1L for materials)[14]. As always we only retain documents with titles and a minimum abstract length of 250 characters, resulting in a control set of 29,952 patents related to agriculture, automotive and materials.

[12] We want to thank Wolfgang Glänzel for his kind help in the development of a search strategy for biotechnology publications.

[13] As all publication of three multidisciplinary journals are included, non-biotechnology publications will also be present as it is not straightforward to isolate biotechnology publications from those multidisciplinary journals.

[14] Patents of the control groups are selected in such a way that there is no overlap with biotechnology patents, i.e., patents classified in both biotechnology IPC classes and one of the control sets IPC classes are not selected for the control groups, only for the biotechnology group. This is of particular interest for the agriculture control group, as this group can be related to biotechnology and share some IPC codes (A01H 1/00 and A01H 4/00).

3.5 Combined dataset

In total, 1,174,021 patent and publication documents are originally selected based on the respective search keys, of which 1,066,632 documents are included in our final setup (all documents having an abstract of substantial length to allow text mining): 88,248 biotechnology patents; 9,952 agriculture patents; 10,000 automotive patents; 10,000 materials patents; 219,713 core biotechnology publications; 647,029 extended biotechnology publications and 81,690 publications from multidisciplinary journals.

4 Derivation of content similarity

4.1 Index parameters and comparison of measures based on Latent Semantic Analysis

We want to match patents and publications based on content similarity, and want to use LSA to derive content similarity from the patent and publication documents. In practice, applying this method involves multiple pre-processing steps to convert a document collection into a document-by-term matrix (tokenization, indexing, weighting, see text mining introduction), and for every of those steps, multiple options are available, resulting in a myriad of choices to arrive at a document-by-term matrix as input for the LSA model. As stated before, the application of LSA in itself also requires a careful choice of the level of dimensionality reduction. Finally, multiple metrics are available to arrive at a similarity value. This heterogeneity in the process to derive content-based similarity measures makes the choice of the best similarity measure (and all corresponding pre-processing options required) very challenging for the purpose at hand, especially as best practices are not readily available. Moreover, our previous experience revealed that common practice does not always yield the best results (see Magerman, Van Looy & Song, 2010).

To shed a light on the feasibility and performance of LSA content-based measures for large scale patent-publication matching, we again compare multiple measures based on multiple weighting options and multiple levels of dimensionality reduction, combined with a limited number of pre-processing steps and the cosine metric. This allows us to

check the effect of weighting and dimensionality reduction options on the performance of the matching method and select the best measure and procedure to arrive at reliable matching results. In total we combine four weighting methods with ten levels of dimensionality reduction, resulting in a setup with 40 measures based on LSA. To complete the assessment of the performance of text mining based measures for patent-publication matching, we also add three measures based on a simple count of the number of common terms between documents.

4.2 Pre-processing choices

The first step in the process it to convert the document collection into a numerical dataset. LSA is based on the Vector Space Model: every document is represented by a vector in a highly dimensional space and every element in the vector represents the weight for a given term for the document at hand.

In practice this is done by an indexer splitting text documents into tokens or terms and compiling a list on how many times a given term appears in a given document. We use Apache-Lucene™, an open source text search engine library, for the indexing[15]. During the indexing process, a minimal number of stop words is removed[16], numbers are removed[17] and stemming is applied (Porter stemmer)[18].

Next we use MathWorks Matlab™, a commercial packet for mathematical and technical computing, for the construction of the vector space by converting the Lucene full text index into a document-by-term matrix[19] [20]. This results in a matrix with 1,066,632 rows (documents) and 729,761 columns (stemmed terms). After removal of terms/stems only appearing in one document, we end up with a document-by-term matrix with 1,066,632 documents and 301,697 terms/stems. This document-by-term matrix contains the raw

[15] http://lucene.apache.org/java/docs/index.html
[16] Based on the Snowball English stop word list (http://snowball.tartarus.org/algorithms/englisch/stop.txt)
[17] Only numbers are removed, i.e. terms that only contain digits. Digits that are part of terms with alphanumerical characters (e.g. chemical formula) are untouched.
[18] http://tartarus.org/~martin/PorterStemmer/index.html
[19] http://www.mathworks.com/products/matlab/
[20] We want to thank Frizo Janssens who was so kind to share his proprietary Matlab code for the import of the full text index into Matlab and compilation of a document-by-term matrix.

term frequencies, i.e. the number of times a given term/stem appears in a given document.

We deliberately choose not to apply more pre-processing tasks, like compound term and collocation detection, because we want to keep the processing simple and automated. These more advanced pre-processing tasks almost always imply more human involvement and manual attention, while we want to opt for an automatic approach.

4.3 4 weighting methods

To improve retrieval and matching performance, raw term frequencies are weighted to take into account the relative importance of a term in a given document or in the complete corpus. Many weighting methods are available, and as in our previous setup we again choose TF-IDF weighting for our current setup as it is commonly used in text mining.

To get a better understanding of the impact of weighting, we again include a non-weighted variant (using the raw term frequencies) and two alternative weighting methods: binary weighting and IDF (inverted document frequency) weighting. In the binary weighting method, we only take the presence or absence of a given term in a given document into account and we ignore the number of occurrences, i.e., the binary weighted frequency of a given term *i* and document *j* is equal to 0 if TF_{ij}=0 and is equal to 1 if $TF_{ij} > 1$ (with TF – term frequency – the number of times a given term appears in a given document). In the IDF weighting method, we combine the binary weighting method with the inverted document frequency, i.e., we replace the raw term frequency for a given term *i* and document j by the IDF value of the given term *i*.

To summarize, we compare four weighting methods for the document-by-term matrix: (1) the raw term frequency (the number of occurrences of the given term in the given document); (2) the binary term frequency (0 if the given term is absent in the given document, 1 if the given term is present in the given document); (3) the inverted document frequency (0 if the given term is absent in the given document, the inverted

document frequency value if the given term is present in given document); and (4) the TF-IDF value (multiplication of term frequency with inverted document frequency).

4.4 10 levels of dimensionality reduction

Dimensionality reduction is an essential part in the LSA method. It truncates the vector space to reveal the underlying or 'latent' semantic structure in the document collection by mapping terms on latent concepts by combining terms in linear relationships. Truncation is done by applying Singular Value Decomposition (SVD) to get a rank-*k* approximation of the original matrix. Dimensionality reduction is supposed to remove the 'noise' due to polysemy and synonymy present in text documents, but the level of dimensionality reduction, or the best selection of the rank (*k*) of the truncated document-by-term matrix, is an open question. As mentioned before, empirical testing shows that the optimal choice for the number of dimensions ranges between 100 and 300 for large datasets. For small datasets, low values of *k* (below 10) seem to work as well (Glenisson, Glänzel et al., 2005), although our previous experience suggests the use of large values of *k*, but also reveals that no dimensionality reduction at all might perform best[21].

In this study, we compare multiple levels of reduction (defined by *k*, the rank order of the truncated document-by-term matrix, i.e. the number of dimensions to retain). We include following nine levels of *k*: 1,000; 500; 300; 200; 100; 50; 25; 10; 5. And to assess the overall value of LSA and dimensionality reduction, we compare these nine levels of dimensionality reduction with a tenth variant, namely no dimensionality reduction at all (which is basically not an LSA-based measure anymore as it is just an application of the cosine metric on the full vector space).

4.5 40 LSA measures and 3 measures based on common terms

To summarize, we compare 40 measures based on LSA by combining 4 levels of term weighting with 9 levels of dimensionality reduction by SVD and no dimensionality

[21] See Magerman, Van Looy & Song, 2010.

reduction at all. For all these 40 measures, we use limited pre-processing options (stop word removal and stemming) and the cosine metric to arrive at a similarity value.

We also include three measures based on the count of the number of terms the patent and publication document have in common. For these three measures, not based on the cosine metric, we use the same pre-processing options as for the 40 measures based on LSA (stop word removal and stemming). To arrive at a similarity metric with values between 0 and 1 starting from the number of common terms, three variants of normalization are used: (1) divide the number of common terms by the minimum of the number of terms of the patent document on the one hand and the number of terms of the publication document on the other hand ('common terms MIN'); (2) divide the number of common terms by the maximum of the number of terms of the patent document and the publication document ('common terms MAX'); and (3) divide the number of common terms by the average of the number of terms of both documents ('common terms AVG')[22]. The second option is more restrictive compared to the first option and only attributes high similarities if both documents are almost identical (the intersection of both documents is equal to the union of both documents: $A + B = A \cap B$). The first option also attributes high similarity if one document is a subset of another document, even if the latter document contains far more information (the intersection of both documents is equal to one of the documents, but potential large remainder or complement of that one document is neglected: $A + B \neq A \cap B$ but $A = A \cap B$). Hence the first option will yield higher similarity values for the same document combinations than the second option, and the third options will be somewhere in-between.

5 Aggregated results

5.1 Similarity calculations

We calculate similarity scores between all 88,248 biotechnology patents and all 948,432 biotechnology publications according to the 43 defined similarity measures. For every

[22] For the ease of reference, we will use 'common terms MIN', 'common terms MAX' and 'common terms AVG' to denote the measures based on the number of common terms and their respective normalization method throughout this document.

patent, the closest 10,000 publications and corresponding similarity scores were retained for every of the 43 measure variants.[23]

We do the same for all 29,952 patents in the control set; for every control patent we calculate similarity scores with all of the 948,432 biotechnology publications according to the 43 defined similarity measure variants and retain again the closest 10,000 publications for every control patent and measure variant.

5.2 Comparison of distributions

To get a first look at the differences amongst measure variants, we compare the distribution of the obtained similarity scores amongst measures. For every measure, we take for every biotechnology patent the closest publication and the corresponding similarity score, hence 88,248 similarity scores for every measure variant. Based on those scores we derive relative distributions displaying the proportion of biotechnology patents having a closest biotechnology publication in a given similarity interval (one histogram for every similarity measure variant). We do the same for all patents related to agriculture, automotive and materials (again one histogram for every similarity measure variant and every control set). For any given measure variant, we can compare the distributions of the similarity scores of the biotechnology patents and the patents related to agriculture, automotive and materials as we used the relative share of patents having a closest publication within a given similarity interval and not the absolute number of patents.

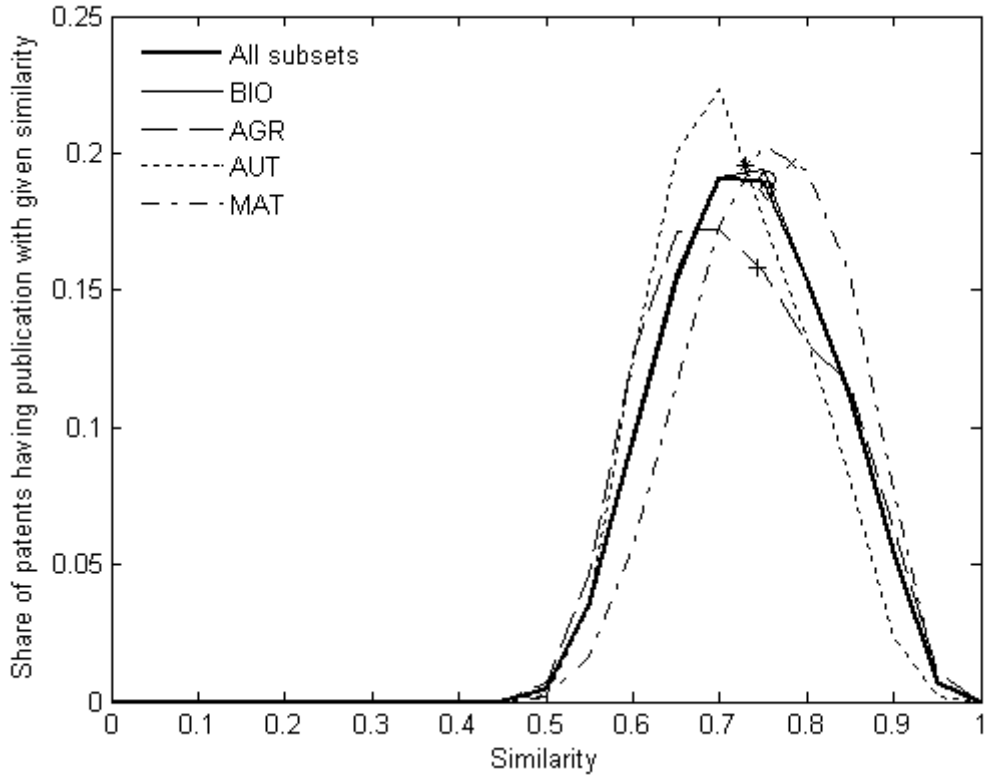
Figure 1 shows the distribution of similarity scores for the patent groups (biotechnology and three control groups) for the similarity measure variant using TF-IDF weighting and SVD of rank 300, a commonly used measure. The Y-axis contains the proportion of patents having a closest publication with similarity given by the X-axis (with intervals of 0.05). Five distributions are combined: one for the biotechnology patents (solid thin line), one for every control group – agriculture (AGR), automotive (AUT) and materials

[23] Retaining all similarities of all 83 billion combinations 43 times is impossible because of current day storage limitations.

(MAT) (non-solid lines) – and one for all patents together – biotechnology patents and all patents from all three control sets (thick solid line).

The distribution of similarity scores of the group of biotechnology patents (solid thin line) falls more or less together with the distribution of all patents (solid thick line) and almost has the same median value. Striking are the relative high similarity scores: the median similarity for all patents is 0.76, or 50% of all patents have a scientific publication with similarity above 0.76. These high average similarity scores are suspicious, although this might simply be a norming or calibration problem.

Figure 1 : Distribution of similarity scores of patents to closest publication according to TF-IDF SVD 300 (markers=median values)



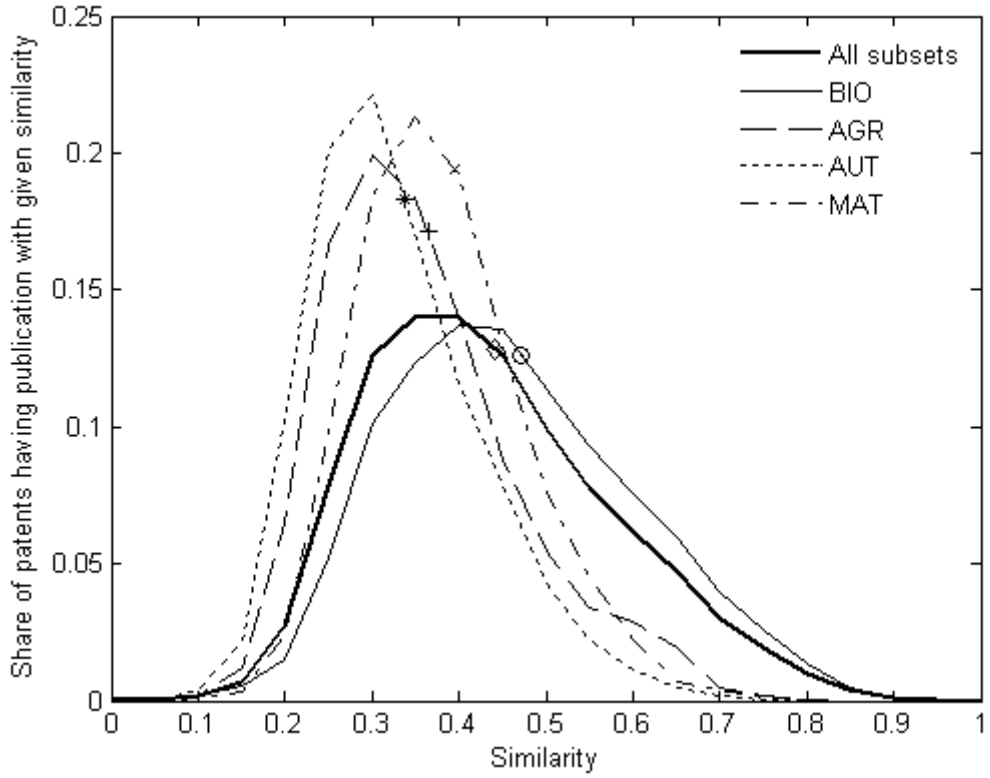
More striking is the distribution of the similarity scores between patents related to materials and their closest biotechnology publication (dot-dash line). We expect the similarity score distributions of control set patents to be to the left of the similarity score distribution of biotechnology patents, as we expect that those control set patents are, on average, less related to biotechnology publications compared to biotechnology patents. However, here we observe that the distribution for materials patents is shifted

to the right compared to the distribution of the group of biotechnology patents and the median value for these materials patents is 0.78. This means that, on average, patents related to materials are more closely related to biotechnology publications than patents related to biotechnology. This is very unlikely and suggests that similarity values based on TF-IDF weighting and SVD of rank 300 do not grasp the real relation between the patent and scientific publication documents.

We observe this phenomenon for all measure variants based on SVD, and the lower the number of retained dimensions, the worse (the more distributions of similarity scores shift to the right and the less difference between the distribution of similarities for patents of the control groups – non-biotechnology patent to biotechnology publication - compared to the group of biotechnology patents). Weighting methods have some effect too: distributions based on binary weighting and IDF weighting are shifted more to the left compared to TF-IDF weighting and raw frequencies, regardless of the number of retained dimensions, and no weighting at all and binary weighting tend to suffer less from the phenomenon of patents of control groups being more similar to scientific biotechnology publications than biotechnology patents. SVD only seems to yield meaningful similarity values when a high number of dimensions are retained (500 or more) and not in combination with TF-IDF weighting (SVD with 1,000 dimensions and TF-IDF weighting still reveals unrealistic distributions).

Figure 2 shows the distribution of similarity scores between patents and their closest biotechnology scientific publication according to the similarity measure variant using TF-IDF weighting without SVD dimensionality reduction. This distribution makes sense: patents from control groups (agriculture, automotive, materials) are on average less similar to biotechnology patents. Even more, there are barely control set patents having high similarity with biotechnology patents. The other weighting methods, combined with no dimensionality reduction, yield similar distributions, although binary and IDF weighting results are slightly more peaked and shifted to the left.

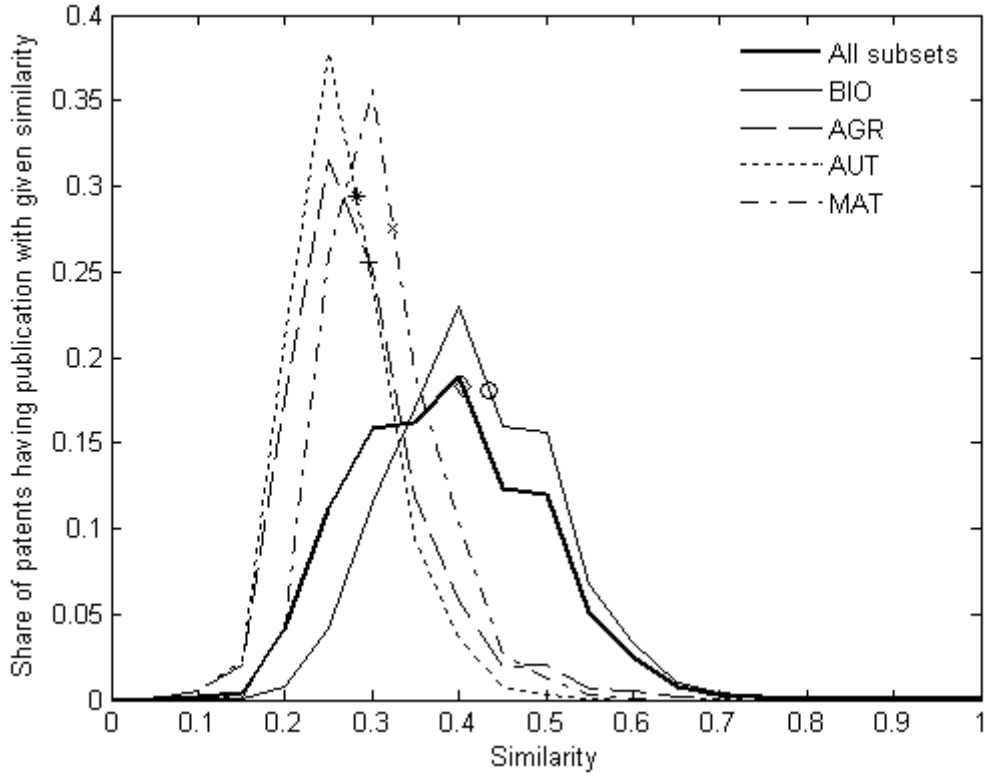
Figure 2 : Distribution of similarity scores of patents to closest publication according to TF-IDF without SVD (markers=median values)



Finally, Figure 3 shows the distribution of similarity scores of the measure variant based on the number of common terms normalized by the minimum of the term length of both documents ('common terms MIN').

Also here we observe an expected distribution with control set patents scoring significantly lower similarity scores compared to the biotechnology patents. All three measures based on the number of common terms yield expected results, although 'common terms MIN' yields the highest distinctive power between biotechnology patents and control set patents. Even more, this measure seems to yield the best distinctive power of all measure variants under study

Figure 3 : Distribution of similarity scores of patents to closest publication according to number of common terms normalized for minimum term length ('common terms MIN') (markers=median values)



5.3 Preliminary conclusions

The comparison of the distribution of the similarity between patents and their closest scientific biotechnology publication and the pattern of biotech patent similarities compared to control patent similarities (agriculture, automotive, materials) amongst measure variants raises questions about the validity of LSA-based measures to match patent documents and scientific publications. Not only do LSA-based measures yield remarkably high similarity scores, they also do not score non-biotechnology patents as less similar to biotechnology publications compared to biotechnology patents, which suggests that these measure variants do not reflect real similarities present in the document collection. The less dimensions are retained, the more obtained similarity scores seem to deviate from the real relations between the documents. This effect is even reinforced when using TF-IDF weighting, a commonly used weighting method. Similarity measures based on the cosine metric without dimensionality reduction seem

to perform better, in combination with any of the tested weighting schemas, as well as the three measures based on the count of common terms. The one normalized by the minimum of the number of terms of both documents ('common terms MIN') seems to yield the best results.

These remarkable results deserve a closer look to patent-publication combinations yielding high similarity values. Table 1 contains the similarity scores of a patent-publication combination scoring high on TF-IDF in combination with SVD (ranging from 0.928 to 0.995, depending on the number of retained dimensions). The title of the patent is: "Process and rotary milking parlor for the identification of a milking stall and an animal, in particular a cow, in a rotary milking parlor." And the title of the scientific publication is: "Growth-behavior of Lactobacillus-acidophilus and biochemical characteristics and acceptability of Acidophilus milk made from camel milk." Title and abstract of both documents make clear that both documents are only (very) slightly related; both are about milk, but the patent is about an apparatus for milking, while the publication is about a comparison of cow milk and camel milk for characteristics on Lactobacillus acidophilus fermentation (see Appendix 2 for the full abstract of both documents). Obtained similarity scores contain considerable variation among weighting methods and dimensionality reduction options.

Table 1 : Similarity scores for patent US7104218 and publication A1994PC04400005 according to various measures

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.511	0.837	0.873	0.905	0.754	0.391	0.368	0.608	0.691	0.673
Binary	0.083	0.057	0.025	0.023	0.056	0.087	-0.030	0.492	0.763	0.750
IDF	0.095	0.168	0.162	0.260	0.375	0.403	0.504	0.532	0.698	0.738
TFIDF	0.364	0.928	0.973	0.986	0.991	0.991	0.995	0.980	0.959	0.960

Especially TF-IDF in combination with SVD yields high scores; other measures yield lower scores, better reflecting the limited relationship between both documents, although all weighting methods yield high values for high levels of dimensionality reduction (low values of *k*, right side of the table). In general, binary and IDF weighting yield lower scores compared to raw frequencies and TF-IDF weighting, although there are some

exceptions. The measures based on the number of common terms yield low scores (0.10, 0.07 and 0.08 for 'common terms MIN', 'common terms MAX' and 'common terms AVG' respectively), in line with the real similarity between the two documents.

Mind also the non-linear relation between similarity scores and dimensionality reduction; lower number of retained dimensions do not necessarily yields the highest similarity scores (see e.g. the results for raw term vectors: starting from 0.837 for 1,000 dimensions it goes up to 0.905 for 300 dimensions, to go down to 0.368 for 50 dimensions to go up again for lower dimensions). This example proves again that the choice of the right level of dimensionality reduction is not straightforward and also that the weighting method has a considerable impact on the results.

6 First validation: comparison of the validity of the measures

6.1 Validation setup

To assess the validity of LSA-based measures and to get more insight in the contribution of weighting and dimensionality reduction levels in the performance of those measures, we set up a validation at the level of individual patent-publication combinations. We select 250 patent-publication cases with variation in similarity scores amongst measure variants. For those 250 cases, we do an independent assessment of experts to rate the similarity on a five-level scale and we check the consistency between the expert assessment and the similarity scores obtained by each of the 43 measure variants for the 250 selected validation cases. This allows us to select the best performing measures.

6.2 Selection of 250 patent-publication combinations to validate

As almost all LSA-based measures tend to attribute high similarity scores to patent-publication combinations, we focus on the selection of patent-publication combinations with high obtained similarity scores to check whether these combinations are indeed similar. At the same time, we want to select patent-publication combinations for validation that have substantial variation in similarity scores amongst measures (it

would not be very informative to select patent-publication combinations scoring high or low on all measures).

Starting point are the 88,248 biotechnology patents. For all those patents, the closest scientific biotechnology publication was selected according to a representative selection of 31 measures (all three measures based on common terms; the non-SVD cosine measure for the four weighting variants; and SVD cosine measure variants with their four weighting variants for *k*=10, 50, 200, 300, 500 and 1,000). For every measure in this selection, the 1,000 most similar patent-publication combinations are retained. After removal of duplicate patent-publication combinations (patent-publication combinations scoring within the top 1,000 for more than one measure), 16,717 patent-publication combinations are left. Out of this selection, 250 combinations were selected in groups of combinations that score high on one measure and low on as many other measures as possible[24] [25].

6.3 Expert assessment of 250 cases

A group of 9 people[26] rated all validation cases (patent-publication combinations) assessing the extent to which the content of the patent document and scientific publication cover the same invention/discovery using a five-level scale: not related at all (1), somewhat related (2), related (3), highly related (4) and identical (5). Every case was independently rated by two people: 176 cases got an identical score by the two raters; 21 cases got scores with a difference of one level; 8 cases got scores with a difference of 2 levels and 45 cases were judged complex. All complex cases and all cases with more than one level difference in scores were assessed by an additional rater resulting in 210

[24] 20 cases scoring high on 'common terms MIN'; 20 cases scoring high on 'common terms MAX'; 10 cases scoring high on 'common terms AVG'; 4x20 cases scoring high on the 4 weighting variants of the cosine measure without dimensionality reduction (one set of 20 cases for each weighting variant); 3x4x10 cases scoring high on the 4 weighting variants of the cosine measure with low (*k*=1000-500), medium (*k*=300-100) and high (*k*=50-5) dimensionality reduction respectively (one set of 10 cases for every weighting variant and dimensionality reduction level).

[25] To compensate for the difference in distributions amongst measures, rank orders were used to evaluate high and low similarities instead of the absolute similarity values.

[26] All nine persons involved in the validation are familiar with patents and publications and IPR and 3 of them are also experts in biotechnology.

cases of total agreement; 39 cases of small disagreement (one level) and 1 remaining case of big disagreement (two levels).

The two independent scores were unified by taking the average of the two scores and rounded to arrive again at a 5 level score. To deal with the potential disagreement amongst raters, two final scores were retained: a conservative one by rounding the average of the scores down to the nearest integer, and an optimistic one by rounding the average up to the nearest integer. Table 2 contains the distribution of similarity levels amongst validated patent-publication combinations according to the conservative and optimistic validation.

Table 2 : Distribution of similarity levels amongst validated patent-publication combinations according to conservative and optimistic validation of experts

Score	Conservative	Optimistic
Identical	161	165
Highly related	8	15
Related	17	10
Somewhat related	10	27
Not related	54	33
Total	250	250

The fact that more than 50% of the cases are judged to be identical has to do with the selection method; our selection started from a set with the 1,000 most similar combinations for each and every measure, as we observe that SVD-based measures tend to attribute high similarity values to patent-publication combinations.

6.4 Check consistency between expert scores and 43 similarity measures

Given the expert assessment of the 250 validation cases, an ANOVA-type of analysis can be used to check the consistency between the expert scores (conservative and optimistic) and the calculated similarity values. Table 3 contains the results of the GLM regression based on 250 patent-publication validation cases for the conservative expert score. This table contains for every measure the R^2 value for the GLM regression with the conservative expert score as independent variable and the similarity values of the

given measure as dependent variable (R^2 values higher than 0.50 are emphasized in bold and italic).

Table 3 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)

Measure		R^2	Measure		R^2
RAW	**No SVD**	***0.61***	TF-IDF	**No SVD**	***0.71***
	SVD 1000	0.34		**SVD 1000**	0.45
	SVD 500	0.31		**SVD 500**	0.34
	SVD 300	0.30		**SVD 300**	0.26
	SVD 200	0.31		**SVD 200**	0.21
	SVD 100	0.30		**SVD 100**	0.17
	SVD 25	0.22		**SVD 25**	0.14
	SVD 5	0.11		**SVD 5**	0.11
BIN	**No SVD**	***0.77***	IDF	**No SVD**	***0.80***
	SVD 1000	***0.65***		**SVD 1000**	***0.63***
	SVD 500	***0.63***		**SVD 500**	***0.57***
	SVD 300	***0.58***		**SVD 300**	***0.54***
	SVD 200	***0.51***		**SVD 200**	***0.51***
	SVD 100	0.45		**SVD 100**	0.49
	SVD 25	0.38		**SVD 25**	0.46
	SVD 5	0.20		**SVD 5**	0.21
Common terms (weighted by min number of terms)					***0.82***
Common terms (weighted by max number of terms)					***0.68***
Common terms (weighted by avg number of terms)					***0.75***

Mean R^2 values in bold denote values higher than 0.5.

Table 3 reveals that the application of SVD dimensionality reduction has a negative impact on the performance of similarity measures: for all weighting methods, dimensionality reduction results in lower R^2 values, i.e. less congruence between the calculated similarity score according to the measure and the similarity level as assessed by the experts. And the larger the dimensionality reduction, the lower the obtained R^2 values. This is especially the case when raw frequencies or TF-IDF weighting is used – remarkable as the combination of TF-IDF weighting and SVD dimensionality reduction is commonly used. Binary and IDF-weighting (lower part of the table) outperforms raw frequencies and TF-IDF-weighting, whether or not SVD is used, and the combination of IDF-weighting without SVD, i.e. a cosine metric based on an IDF-weighted document-by-term matrix, yields the highest R^2 (0.80) of all cosine-based measures. Striking is also that simple metrics based on the number of common terms score very high, even more,

the metric based on the number of common terms weighted by the minimum number of terms of both documents ('common terms MIN') yields the highest R^2 value (0.82).

When the optimistic expert scores are used instead of the conservative expert scores, results stay the same: despite small changes in R^2 (upward for some measures, downwards for others), conclusions about SVD dimensionality reduction, weighting method and best measures remain the same.

Also, when we convert the 5-level scale expert scores to 2-level scale expert scores (identical versus not-identical) to focus on the identification of patent-publication pairs, results stay the same.

6.5 First validation results

Our ANOVA results reveal that the similarity measure 'common terms MIN' best matches our expert validation. Of course it does not come to a complete surprise that measures based on the number of common terms perform that well: the more terms in common, the more you can expect both documents to be similar. But on the other hand these simple measures based on the number of common terms might miss relevant matches because they do not deal with language related issues like homonymy, polysemy and synonymy. It is remarkable that despite this lack of complexity these measures come closest to the expert assessment of similarity – clearly beating LSA measures that do claim to deal with typical language issues. Another remarkable observation is the consistency between 'common terms MIN' and the presence or absence of a publication author in the list of patent inventors – a strong indication whether or not the patent and publication is identical, i.e. shares the same contents (methodology, findings, discovery). All patent-publication combinations with a similarity of 0.59 or above according to this measure do have a publication author listed as patent inventor, and all combinations with a similarity of 0.50 or below do not have a publication author listed as patent inventor (with one exception with a similarity value

of 0.16). In between are 5 cases, 3 with and 2 without a publication author listed as patent inventor. This consistency is a strong indication of the validity of this measure[27].

If we take 0.55 as a threshold value (in between the zone with shared inventor/author and the zone without shared inventor/author) and translate the 5-scale expert score to a 2-scale score (identical or not identical – as we are primarily concerned about finding patent-publication pairs, hence primarily concerned about identical versus not-identical patent-publication combinations) we obtain a confusion matrix as displayed in Table 4 (using the conservative expert scores).

Table 4 : Confusion matrix for the measure based on the number of common terms weighted by minimum number of terms of both documents (based on conservative expert scores of 250 validation cases with threshold value of 0.55)

			Measure COMMON TERMS MIN	
			Identical	Not identical
			168	82
Expert opinion	Identical	161	160	1
	Not identical	89	8	81

This results in a precision of 0.95 (percentage of document combinations classified as related by the automated method that are correct according to the experts: 160/168 – to what extent is the automated method correct when it predicts a match) and a recall of 0.99 (percentage of document combinations that are related according to the experts that are classified as related by the automated method: 160/161 – to what extent does the automated method not miss relevant matches). When the optimistic expert scores are used, the number of identical patent-publication combinations according to the

[27] Although the presence or absence of a shared inventor/author is a strong additional indication of content similarity, using this criterion on its own to identify patent-publication combinations is not straightforward – as stated in the introduction – because of practical reasons (how to deal with spelling errors in names; presence or absence of initials and middle names; homonyms) and conceptual reasons (the same person can be involved in multiple discoveries/inventions, hence two documents of the same inventor/author can have a complete different contents). It is the combination of content relatedness and presence of shared inventor/author that yields a robust indicator.

experts raise from 161 to 165, and this results in an even higher precision of 0.98 and recall of 0.99.

Although the validation seems to result in excellent precision and recall scores for an automated classification method, these scores are misleading because the distribution of the obtained similarity scores according to measure 'common terms MIN' is completely different within the validation sample and the total population. We have an underrepresentation of document combinations that are less related in our validation sample because the selection of validation cases was primarily based on combinations scoring high on at least one measure. As listed in Table 2, more than 65% of the cases in the validation sample were rated identical by the experts, while the relative number of patent-publication pairs in the full populations will be far, far lower. In reality, the number of document cases with average or low similarity scores will completely outnumber the cases with high scores. And although the relative number of misclassifications might by reasonable, this large group with average scores will result in a high number of mismatches in absolute terms, pulling down the relative number of correct classifications and negatively influencing precision and recall. To get reliable precision and recall scores, the relative number of mismatches has to be combined with the absolute number of document combinations to correct for the differences in distribution. We will come back on this issue later on when presenting validation results based on an extended validation set.

7 Additional validation: validation based on control sets

7.1 Validation setup

Apart from the expert validation, the control sets can be used for additional validation. As described earlier, three control sets were created with patents related to agriculture, automotive and materials, with 29,952 patents in total. These patents are presumed to be unrelated to biotechnology publications, meaning that we do not expect to find biotechnology publications having a high content similarity with any of these control set

patents.[28] If we apply our measure 'common terms MIN' on all combinations of the 29,952 control set patents and the 948,432 biotechnology publications, we expect not to find high similarity scores.

7.2 Additional validation results

In total we find 126 combinations of control set patents and biotechnology publications with a similarity value of 0.60 or above according to the measure 'common terms MIN' (about 0.04% of all control set patents), compared to 4,499 combinations of biotechnology patents and biotechnology publications (about 5.10% of all biotechnology patents). This significant difference in the ratio of patent-publication combinations with high content similarity between the group of biotechnology patents and the group of control patents is again an indication of the validity of our measure. Yet it might be interesting to dig into those 126 control set cases with high similarity. 51 of those cases have a similarity of 0.70 and above, and 12 cases even have similarities of 0.80 and above.

Appendix 3 contains an example of a combination of a control set patent and biotechnology publication (common terms min = 0.82; common terms max = 0.06). This example demonstrates the weakness of using the minimum number of terms of both documents as weight to normalize the number of common terms to arrive at a metric. The patent abstract is far shorter compared to the publication abstract, and as almost all terms of the patent abstract are present in the publication abstract, a high similarity is obtained when using the minimum number of terms of both documents as weighting factor. This approach seems to make sense in general; if the abstract of one document is a subset of the abstract of the other document, they can be regarded as identical. We checked this for multiple cases where there is a big difference in the similarity value based on measure 'common terms MIN' and measure 'common terms MAX' – an indication of document combinations with unbalanced text length – and indeed, for the

[28] As stated before, patents of the control groups are selected in such a way that there is no overlap with biotechnology patents, i.e., patents classified in both biotechnology IPC classes and one of the control set IPC classes are not selected for the control groups, only for the biotechnology group. This is of particular interest for the agriculture control group, as this group can be related to biotechnology and share some IPC codes (A01H 1/00 and A01H 4/00).

vast majority of those cases the longer document just contains more details or a longer introduction or results, but the actual relevant contents is the same. So there is some anecdotic evidence to back up the use of the minimum number of terms as weight (on top of the empirical results of the ANOVA-analysis revealing this measure as the best performing one). However, when one of the documents is too small, or when the difference in length is too big, using the minimum number of terms as weight leads to unreliable results (even for human experts it becomes difficult to assess similarity for these cases).

If we go back to our 126 control set patents with high similarity with a biotechnology publication (weighted by the minimum number of terms of both documents – measure 'common terms MIN'), it is striking that all of them do have low similarity values when the maximum number of terms of both documents is used as weight (measure 'common terms MAX') - i.e. there is a big difference in the length of both documents. Only 21 of those cases have a similarity 'common terms MAX' above 0.10, and only 2 above 0.20 (with a maximum of 0.25). In our validation set of 250 cases, 71 cases have a similarity 'common terms MAX' of 0.25 or below; and only 2 of those cases are rated as identical by the experts (one cases of 0.24 and one case of 0.20).

7.3 Additional criterion

The insights of the additional validation suggest that a correction is needed for document combinations with one small and one large document. For those cases, our best performing measure 'common terms MIN' might be misleading and an additional criterion based on document length is needed. Instead of adding an absolute criterion based on document size, we examine the impact of an additional relative measure, as we have already one measure available: measure 'common terms MAX'. So we combine the primary criterion based on measure 'common terms MIN' (e.g. above 0.55) with a secondary criterion based on measure 'common terms MAX' to correct for doubt cases. The results of the additional validation based on the control sets suggests that the threshold for this secondary criterion 'common terms MAX' is somewhere between 0.20 and 0.30 (almost all combinations from the control set score below 0.20 on this

criterion with a few exceptions between 0.20 and 0.30, and all combinations in our validation sample of 250 expert rated cases scoring below 0.20 on this criterion are rated not identical by the experts).

Applying this secondary criterion 'common terms MAX' with threshold around 0.20 does not influence the classification of the 250 expert rated cases in our validation sample because none of those cases with primary criterion 'common terms MIN' above 0.55 score below 0.20 on the secondary criterion (4 cases score between 0.20 and 0.30, and all four are rated identical by the experts).

However, the control set validation proves that setting the threshold value for the secondary criterion 'common terms MAX' does has a significant impact (e.g. setting the value to 0.20 would discard all matches found for the control set patents). If we look at the global biotechnology dataset (88,248 biotechnology patents and 948,432 biotechnology publications), and take the 1,000 closest combinations for every patent according to measure 'common terms MIN', there are 112,847 patent-publication combinations above 0.55 for the primary criterion 'common terms MIN', but the vast majority of those combinations score low on the secondary criterion 'common terms MAX'. Table 5 contains the distribution of the 'common terms MAX' scores for the combinations above 0.55 for 'common terms MIN'.

Table 5 : Distribution of second criterion scores 'common terms MAX' for all patent-publication combinations with primary criterion 'common terms MIN' above 0.55

Second criterion 'COMMON TERMS MAX' range	Number of patent-publication combinations	Number of patent-publication combinations (cumulative)
$0.35 \leq x$	631	631
$0.30 \leq x < 0.35$	262	893
$0.25 \leq x < 0.30$	747	1,640
$0.20 \leq x < 0.25$	2,856	4,496
$0.15 \leq x < 0.20$	14,053	18,549
$0.10 \leq x < 0.15$	56,093	74,642
$0 \leq x < 0.10$	38,205	112,847

The figures in table Table 5 make clear that matching results are extremely sensitive to threshold setting; even within the range of 0.20-0.30 the impact on the number of

matches is significant (from 4,496 combinations labelled as identical by the automatic method for a threshold value of 0.55 for 'common terms MIN' and 0.20 for 'common terms MAX' to 893 combinations labelled as identical for the same threshold value for 'common terms MIN' but a threshold value of 0.30 for 'common terms MAX').

8 Final validation: selection of 50 additional cases for expert validation

8.1 Validation setup

As the first validation set of 250 cases for validation cases does not allow for careful selection of the threshold value for the secondary criterion - as we do not have enough cases with low scores on 'common terms MAX' in our validation sample - 50 additional cases were selected. For the selection of these additional cases, we do not only look for cases with low scores on the secondary criterion 'common terms MAX', but also for potential false negatives and false positives for the primary criterion 'common terms MIN'. The idea is the create a robustness check for a classification method based on 'common terms MIN' and 'common terms MAX' by deliberately selecting additional validation cases that are though, i.e. difficult to classify because they are in the grey zone between identical and not-identical combinations or cases that are expected to be misclassified based on the information we have. To obtain a balanced and representative selection, we split up the selection of additional validation cases by similarity range for the primary criterion 'common terms MIN':

0.81-1.00 : this is normally the save zone were we only expect to find patent-publication combinations that are identical (the first validation only revealed one case not rated as 'identical' by the experts). For this zone, we are interested in potential false positives introduced by the primary criterion, so we select 5 cases without shared inventor/author because these are unlikely to be identical (all those cases happen to have a low score on the secondary criterion).

0.71-0.80 : this is still rather a save zone (the first validation only revealed one or two cases not rated as 'identical' by the experts - depending whether conservative or

optimistic expert scores are used). For this zone, we are interested in potential false positives introduced by the primary criterion and in potential false negatives introduced by the secondary criterion. We take 5 cases without shared inventor/author (potential false positives) and with high scores on the secondary criterion (how to set second criterion threshold to discard false positives); 5 cases with shared inventor/author and low scores on the secondary criterion (to what extend will the secondary criterion introduce false negatives); and 10 cases with shared inventor/author and a secondary criterion value around 0.3 (to help finding a solid threshold for the secondary criterion), so 20 cases in total.

0.61-0.70 : this is a grey zone with multiple mismatches according to the first validation. We follow the same logic for the selection of cases as for the previous range: 8 cases without shared inventor/author and with high scores on the secondary criterion and 7 cases with shared inventor/author and low scores on the secondary criterion, so 15 cases in total.

Below 0.61 : here we are interested in false positives in the frontier zone ('common terms MIN' in the range of 0.55-0.60) and in false negatives for lower values on the primary criterion 'common terms MIN'. We select 5 cases scoring high on the primary criterion (within this range) and without shared inventor/author or low value on the secondary criterion, and 5 cases scoring low on the primary criterion (within this range) and with shared inventor/author.

8.2 Final validation results

Those 50 cases were again rated by two experts as in the first validation. Table 6 contains the result of the validation for every range and subset based on conservative expert scores.

Table 6 : Expert validation results (conservative) for 50 additional cases by primary criterion range ('common terms MIN') and validation subset

Range primary criterion	Validation subset	Total cases	IDENTICAL ACCORDING TO EXPERTS		NOT IDENTICAL ACCORDING TO EXPERTS	
			Cases	Range secondary criterion	Cases	Range secondary criterion
0.81-1.00	Potential false positives	5	0		5	0.11-0.18
0.71-0.80	Potential false positives	5	1	0.28	4	0.25-0.33
0.71-0.80	Potential false negatives	5	0		5	0.10-0.16
0.71-0.80	Secondary criterion around 0.3	10	9	0.31-0.41	1	0.36
0.61-0.70	Potential false positives	8	1	0.32	7	0.29-0.45
0.61-0.70	Potential false negatives	7	2	0.24-0.26	5	0.20-0.26
< 0.61	Potential false positives	5	1	0.35	4	0.30-0.51
< 0.61	Potential false negatives	5	2	0.36	3	0.35-0.40

For the first range ('common terms MIN' in the range of 0.81-1.00) results are good for the false positives: all cases that score high on 'common terms MIN' but that were suspect of being false positives (because they had no shared inventor/author) are rated as not identical by the expert validation. It is clear that the secondary criterion based on 'common terms MAX' easily discards all those false positives with a clear threshold value of 0.18).

For the second range (0.71-0.80), results are still reasonable, although a proper selection of the threshold value for the secondary criterion is not clear. A threshold value below 0.36 will introduce false positives, but a threshold value above 0.28 will introduce false negatives, so no clear cut-off point exists and a trade-off has to be made (e.g. a threshold value of 0.30 results in 3 false positives and 1 false negative).

For the third range (0.61-0.70), the overlap gets bigger and the choice for a threshold value for the secondary criterion gets complicated. A threshold value below 0.45 will introduce false positives, but a threshold value above 0.24 will introduce false negatives, so again no clear cut-off point exists and a trade-off has to be made (e.g. a threshold value of 0.30 results in 2 false negatives and 5 false positive).

Finally, the last range (< 0.61) requires even higher values for the secondary criterion to discard false positives (0.51) but the overlap is not much bigger compared to the previous range (at least 0.36 to prevent false negatives). We observe identical combinations (according to the expert validation) up to a similarity of 0.46 for 'common terms MIN', but distinction from false positives is difficult, even with 'common terms MAX' as secondary criterion.

Based on these results, it makes sense to add a secondary criterion based on the number of common terms weighted by the maximum number of terms of documents to eliminate potential false positives with minimal introduction of false negatives. But the results in Table 6 also reveal that for lower values of 'common terms MIN' a clear distinction between identical and non-identical combinations is not possible.

It is clear that setting thresholds on the primary criterion ('common terms MIN') and secondary criterion ('common terms MAX') is a trade-off between false positives and false negatives, or precision and recall.

Table 7 contains precision and recall for different thresholds on the primary and secondary criteria (optimal precision, optimal recall, and balanced precision/recall) based on all 300 cases rated by experts (both for the conservative and optimistic expert scores).

Table 7 : Precision and recall for different thresholds on primary and secondary criterion (optimal precision, optimal recall, balanced precision) (based on conservative and optimistic expert scores for 300 validated cases)

Primary criterion	Secondary criterion	CONSERVATIVE EXPERT OPINION		OPTIMISTIC EXPERT OPINION	
		Precision	Recall	Precision	Recall
0.50	0.10	0.81	0.99	0.88	0.98
0.50	0.32	0.91	0.92	0.94	0.88
0.50	0.61	0.98	0.55	1.00	0.51
0.55	0.10	0.82	0.98	0.88	0.97
0.55	0.30	0.90	0.93	0.93	0.89
0.55	0.61	0.98	0.55	1.00	0.51
0.60	0.10	0.83	0.95	0.98	0.94
0.60	0.29	0.91	0.92	0.94	0.88
0.60	0.61	0.98	0.55	1.00	0.51

Optimal precision scores can be obtained with a recall around 0.55/0.51, optimal recall scores can be obtained with a precision around 0.81/0.88 and balanced precision/recall scores around 0.90 are possible for both precision and recall at the same time (e.g. 'common terms MIN' above 0.55 and 'common terms MAX' above 0.30).

As stated before, we have to keep in mind that the precision and recall figures listed in Table 7 are not representative for the total population because obtained similarity values for 'common terms MIN' and 'common terms MAX' are not equally distributed in the validation sample and the total population (very high number of identical document combinations in the validation sample). As there are only a very limited amount of document combinations scoring high on the proposed measures in the total population, it is more appropriate to derive precision and recall measures based on validation cases scoring around the threshold values. Take for instance a threshold value of 0.55 for 'common terms MIN' and 0.30 for 'common terms MAX'. According to the conservative expert validation, this would result in a precision of 0.90 (184 combinations classified as pair by the automated method in the validation set of which 165 are real pairs according to the experts) and a recall of 0.93 (165 real pairs retrieved by the automated method in the validation set compared to 177 real pairs identified by the experts). Applied on the full population this would mean that we would label 893 patent publication combinations as identical (see Table 5), and by doing so, about 89 of those cases would be wrongly labelled identical (10%), and at the same time we would miss about 61 cases (7%). However, if we look at the cases with 'common terms MAX' between 0.30 and 0.25 in our validation set, we find 9 cases of which 5 cases are assessed as identical by the experts, or 55%. These matches will be missed by the automated method when the threshold for the secondary criterion is set to 0.30. If this 55% match rate is representative for the whole population in the range of 0.25 and 0.30 for 'common terms MAX', we would miss 411 patent-publication pairs in this range for 'common terms MAX'[29]. In the range of 0.20-0.25 for 'common terms MAX', we observe 34% matches in the validation sample. Again according to Table 5 we would miss an

[29] According to table Table 5, we have 747 cases in the total population with 'common terms MIN' above 0.55 and 'common terms MAX' in the range of 0.25 and 0.30, of which 55% or 411 cases are expected to be identical according to the validation.

additional 971 patent-publication pairs. If we continue this reasoning, we and up with a match rate of 20% for the range 0.15-0.20 resulting in another 2,811 missed patent publication pairs. This makes a total of 4,193 expected missed patent-publication pairs, far more than the 61 cases we initially expected. According to these estimations, the real recall rate is 16%.

The same problem occurs for precision rates. In the range 0.30-0.35 for 'common terms MAX', the precision rate in the validation set 53%. Again according to table Table 5 this would mean already 123 false positives. For 'common terms MIN' equal to 0.35 and above, precision rate is 94% hence 38 additional false positives, or 161 false positives in total, again more than the 89 false positives initially expected according to the precision rates in table Table 7. According to the estimations based on the full dataset, the real precision rate is 82%.

The button line is that precision and recall rates derived from the validation sample are not representative for the whole population because we have far, far more patent-publication combinations scoring low on the proposed distance measures while we initially calculated precision and recall rates from sample data with an overrepresentation of patent-publication combinations scoring high on the respective distance measures. Especially recall rates are suffering from this issue. However, the magnitude of the difference between the precision and recall rates averaged over the validation sample and the real rates based on the distribution in the global population heavily depends on the representativeness of those cases scoring low in the validation set. As the previous examples describes, these derived numbers are based on only 28 cases scoring less than 0.30 for 'common terms MAX' (given a score of 0.55 or above on 'common terms MIN'). More validation cases with lower scores are needed to get a more reliable estimate of the real precision and recall. But to be at the save side, the threshold on 'common terms MIN' has to be increased to get acceptable precision rates (e.g. to 0.60).

Precision can be improved by introducing a third criterion: the presence of a shared inventor/author. Although this extra criterion helps to make results more robust, large scale application on big datasets might not be straightforward.

9 Where does it go wrong for TF-IDF and SVD

9.1 Weighting issues

The most remarkable finding of this study is the bad performance of SVD-based measures, even with commonly used pre-processing options and levels of dimensionality reduction (e.g. TF-IDF weighting in combination with SVD with 300-1,000 dimensions).

When it comes to the influence of weighting, Table 3 reveals that weighting methods taking into account term frequencies (raw frequencies and TF-IDF weighting) perform worse compared to weighting methods ignoring term frequencies for all levels of dimensionality reduction. In line with these findings we also observe better performance for the measures based on the number of common terms, measures which also ignore term frequencies.

Looking at individual cases gives some insight in the implications of the choice of a weighting method. In general, including term frequencies is expected to generate better results as the number of times a given term appears in one document is an indication of the importance of that term in that particular document. However, for our patent-publication document combinations (mostly of a rather moderate length and with highly technical content), this additional notion of importance derived from term frequencies seems to be of less relevance in the assessment of similarity of the documents. Indeed, when looking at multiple document combinations, the human judgement on similarity is far more driven by the kind of terms in the documents rather than the number of times a particular term appears in a document. This observation explains why weighting methods taking into account term frequencies do not perform better, but not why they perform worse. Again looking at individual cases reveals some additional insights.

First of all, stemming errors and tokenization and parsing issues sometimes cause artificial inflation of term frequencies, magnifying the impact of the underlying stemming and tokenization errors.

Appendix 4 contains an example of a patent-publication combination where the amplification of a stemming error results in misleading similarity scores for weighting methods taking into account term frequency. The patent document is about an incubator with external gas feed. The publication document is about gibberellin metabolism in suspension-cultured cells of raphanus-sativus. Both documents have nothing in common, yet score high on some measures (and score significantly higher for measures including term frequencies). Both documents have only two (stemmed) terms in common, 'feed' and 'ga'. But the stemmed term 'ga' occurs 9 times in the patent document and 29 times in the publication document, resulting in high weights when the term frequency is included. But the stemmed term 'ga' in the patent document is a stemming error derived from 'gas', while the stemmed term 'ga' in the publication document is an abbreviation of 'gibberellin' and has nothing to do with the stemmed term 'ga' in the patent document. For weighting methods not taking term frequency into account, this stemming error counts as just one (be it wrongly) matching term, but for weighting methods using term frequency, this stemming error is magnified and leads to erroneous results.

Appendix 5 contains an example of a patent-publication combination where tokenization and parsing issues result in misleading similarity values for weighting methods taking into account term frequencies. Again both documents are not related and have only two terms in common: 'alpha' and 'beta'. Both of these terms occur a lot in both documents as part of chemical formulas, and these high term frequencies result in higher similarity values for weighting methods based on term frequencies. But the larger chemical formulas these terms are part of, are not related. It would probably be better to parse and index those formulas as one piece, but this is not straightforward.

Secondly, we observe that words with a particular meaning and hence very relevant in the assessment of similarity tend to have smaller term frequencies compared to natural language words with a more general meaning. For weighting methods including term frequencies, the weight of these more general natural language words becomes too influential in the derivation of similarity by the cosine metric. This issue might be specific to the technical nature of the documents – i.e. for our set of patent and

publication documents, low frequency technical words are far more important for the assessment of similarity compared to higher frequency natural language words. Weighting terms by their respective IDF values does only partially correct this problem; TF-IDF performs better than no weighting at all, and IDF performs (slightly) better than binary weighting, but TF-IDF still performs worse compared to binary or IDF weighting.

Given these insights, it might be worthwhile to investigate to leave stemming out of the pre-processing steps and to devote additional efforts for a more advanced tokenization and parsing, especially to better deal with chemical formula. Another approach is to improve feature selection to eliminate or further down-weight terms which are too general in meaning to be significant in the derivation of similarity.

9.2 SVD issues

It is not clear why LSA – or SVD – fails, or why SVD tends to assign unrealistic high similarity scores to document combinations - mind in that respect the high similarity scores for the patents in the materials control set. While some anecdotic evidence exists to explain differences in weighting performance, disentangling the bad performance of SVD in general is of a different level of complexity. Looking at individual cases is not very informative as it is virtually impossible to trace back term vectors after SVD to the original terms and contents. The document-by-concept matrix compiled by the SVD solution contains the scores of all documents on newly formed latent concepts, and every latent concept consists of a linear combination of all original terms, i.e. a linear combination with 301,697 components.

There are however some general reasons why LSA or SVD might fail for our dataset. The first reason is that the dataset might not be large enough to derive the latent structure. This is however very unlikely for our dataset as it contains almost one million documents. A related issue might be that the individual documents are not long enough to grasp the contents of the documents. This issue might be relevant for our dataset as we work with titles and abstracts, and especially patent abstracts tend to be rather

small (about 39 unique terms on average for patents and 65 unique terms on average for publications[30]). We will come back to this issue later on.

Another reason for the unfulfilled expectations might be that the chosen levels of dimensionality reduction are not appropriate for our dataset, or that our derived SVD solutions for our selection of *k*-values accidentally do not grasp the latent structure of the data. The former deserves more attention, although literature suggests 300 to 1,000 concepts is enough to capture the topics in a document set, which is the range we included in our setup. We will also come back to this issue later on. The latter is very unlikely: SVD solutions are based on the singular values of the full document-by-term matrix. Changing the number of retained dimensions/concepts does not alter the values of the singular values, it only alters the number of singular values and singular vectors taken into account to approximate the original document-by-term matrix. As singular values are ordered by magnitude and values drop significantly, small changes in the number of retained dimensions cannot have big effects once beyond the first tens or hundreds of singular values. Moreover, we have four fundamentally different SVD derivations because of the four weighting methods, and all those variants yield SVD based measures that underperform compared to cosine measures on the full vector space.

A complete different kind of issues resides in the technical nature of the documents. Maybe the specific context of patent and publication documents does not allow the method to achieve its full potential. LSA is intended to derive meaning from text based on large samples of 'narrative' documents. The distinctive language use within our dataset might not be appropriate (might especially be of a concern for patent documents where phrasing might reflect tactical and strategic consideration more than technical disclosure, e.g. to maximize legal claims to get broad application protection or to disguise the real contents to mislead competitors). A related issue might be that we are combining patent documents with scientific publications, two document spheres that might be too different to derive a latent structure that fits both. These potential causes of failure might look farfetched, but it is clear when reading patent abstracts that

[30] After stop word removal, stemming and removal of words appearing in only one document.

such documents have little in common with typical applications as e.g. the ones described in the Handbook of Latent Semantic Analysis (Landauer, McNamara et al., 2007). However, finding evidence for these raised issues is not straightforward. One could set up validation exercises as the ones deployed in this study to evaluate LSA performance on a distinct subset with only patents and a distinct subset with only scientific publications. Another avenue might be to compare the LSA performance when more descriptive abstract are used, e.g. using the *Derwent Abstracts* as available in the *Derwent World Patent Index* (Thomson Reuters Derwent World Patent Index), which are abstracts rewritten by scientifically-trained editors detailing claims and disclosures of the invention and highlighting main use and advantages. Pursuing such additional research efforts might reveal interesting information on the applicability of LSA on patent and publication data, but goes beyond the limitations of this study.

One final reason why LSA might fall short is the limitation to Euclidean geometry as imposed by the assumption of LSA that documents are represented as vectors in a vector space. In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of point in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007). This issue however is a general one and not directly related to the limitations of our patent and publication dataset; it is related to problems when dealing with high-dimensional spaces ('curse of dimensionality'). In high dimensional spaces all data appear to be sparse and dissimilar, preventing efficient identification of communalities. Other text mining techniques not relying on spatial representations might be more appropriate, like generative topic models as Probabilistic Latent Semantic Modelling (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), but the exploration of those methods are again beyond the limits of this study.

In the next parts, we will elaborate more on the impact of document size and the impact of the number of retained dimensions/concept on the performance of SVD.

9.3 Impact of document size on SVD performance

The poor performance of SVD might be related to the document size, as especially patent abstracts tend to be short. To get more insight in this issue, we include document size when we check for the congruence between obtained calculated similarity scores and the expert validation scores. For all patent-publication combinations in the validation sample, we use the minimum document size, i.e. the minimum of the number of terms of the patent document and the publication document, as indicator of the document length. If we include this in the regression analysis, i.e. if we take again the compiled similarity measure variants as dependent variables and we take the expert score and the document size as independent variables, results reveal that document size has no impact on the similarity scores for our measure 'common terms MIN', but that the impact is significant (at the 5% level) for all cosine based measures without SVD. For binary and IDF weighting in combination with SVD, document size also has a significant impact on the similarity score, but not for SVD in combination with TF-IDF weighting or raw frequencies. Overall, the impact is small, except for binary and IDF weighting in combination with SVD, where R^2 values can improve with 8 to 11 percentage points compared to the model with only the expert score as independent variable. Whether we use the total number of terms or the number of distinct terms does not make a lot difference, although results are somewhat softened when the number of distinct terms is used to derive the document size indicator.

Likewise, we also had a look at the difference in document size within a patent-publication document, as we know there are many combinations with a small document combined with a large document. Now we used the ratio between the number of terms of the smallest document and the number of terms of the largest document as indicator of document size difference. If we take the expert scores and the document size difference as independent variables, we see comparable results as for the document size effect, but with stronger impact. Again the impact is not significant for our measure 'common terms MIN' but is significant for all cosine based measures without SVD, for binary weighting and IDF weighting with SVD, and for TF-IDF weighting and raw

frequencies with SVD 1,000. The impact of the document size difference is higher than the impact of the document size, with R^2 values increasing with 15 to 20 percentage points for binary and IDF weighting with SVD.

If we also include both document size and document size difference in the model, and the interaction between document size and document size difference, we observe that the significance of the document size disappears and the impact of document size difference remains.

These results tend to suggest that SVD based measures in combination with binary and IDF weighting are influenced by the document size difference, which might be an explanation for the poor performance. However, this is not a complete explanation as this impact is rather moderate for SVD in combination with TF-IDF weighting and no weighting at all, while those measures perform worst.

In a last analysis trying to disentangle the relation between document size and performance, we looked at direct influence of document size and measure performance. All patent-publication pairs in the validation sample where uniformly divided into three groups: group one with small documents (measured as before by the minimum number of terms of the patent and publication document); group two with medium size documents; and group three with large documents. Now we perform a regression analysis with the similarity measures as dependent variable and the expert scores as independent variable for each of the three groups. For measure 'common terms MIN', performance goes down for larger documents (R^2 of 87% for small documents to 61% for large documents). For cosine based measures without SVD, performance of binary and IDF weighting also goes down by about 15 percentage points; for raw frequencies performance goes up considerable (R^2 from 38% to 72%), and for TF-IDF weighting performance remains more or less constant (R^2 around 64%). For SVD with low levels of dimensionality reduction (k=1,000) we see the performance slightly going down for larger documents for binary and IDF weighting, but heavily going up for TF-IDF weighting and no weighting. If we do the same kind of analysis with document size difference instead of document size, we observe a considerable increase in performance of SVD with low levels of dimensionality reduction (k=1,000) for TF-IDF

weighting and raw frequencies for patent-publication combinations that are balanced in document size, while performance remains more or less constant for binary and IDF weighting.

To summarize, we see that document size and document size difference has a different impact on the measures depending on the weighting schema used. The three measures based on the number of common terms, and the cosine measures without SVD and without term frequencies (binary and IDF weighting) tend to perform worse for larger documents and/or documents of equal size. There seems to be a normalization problem for these measures. More important is that SVD-based measures and especially TF-IDF measures in combination with low levels of dimensionality reduction perform far better for larger and more balanced patent-publication combinations, although not yet beating our preferred measures 'common terms MIN'. In this respect it would be interesting to combine both document size and document size difference in the same analysis. The problem is that the number of observations in the validation set becomes low for some combinations of document size and document size difference, and that the variance amongst expert scores becomes very low for some of these combinations.

We have to be careful in deriving hard evidence from these analyses, as the validation sample is rather small - especially when split up by size and size difference - but it seems that document size and document size difference have an impact on SVD based measures and might at least partially explain the bad performance of e.g. TF-IDF weighting in combination with SVD because of our rather short patent abstract documents. A larger validation sample with a more balanced design when it comes to document size and document size difference is required to disentangle this further.

9.4 Impact of the number of retained dimensions/concepts and stability of the SVD solution

As stated before, the choice of *k*, the number of retained dimensions or concepts, is not straightforward. In this study, the maximum value of *k* taken into consideration was 1,000 because of computational limitations. Although literature suggest to take 100 to 300 concepts, the variety of topics present in our patent and publication set might

require more concept to be taken into account to grasp the latent structure of the dataset. Computationally limitations prevent us from deriving SVD solution with more than 1,000 retained dimensions/concepts for our large dataset, but using a smaller sample allows us to go beyond 1,000 retained concepts in the SVD calculation.

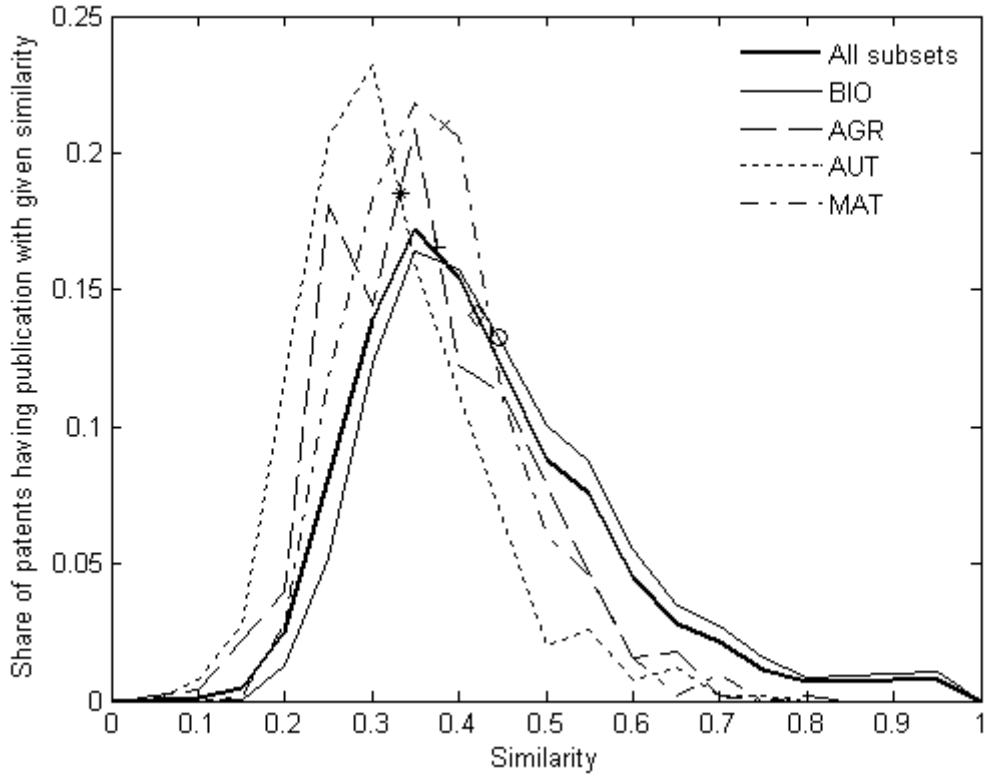
We started from the original raw document-by-term matrix, the document-by-term matrix after IDF weighting, and the document-by-term matrix after TF-IDF weighting. Every time we selected 5% random patents from each group (biotechnology, agriculture, automotive and materials) and 5% random publications from our original dataset. However, we made sure that patent-publication combinations that are present in our expert validation set are also present in all samples. This resulted in three different subsamples of 53,332 patent and publication documents, based on three different weighting methods (mind that not only the weighting method is different, but that selected patents and publications are also different, except for the patent-publication combinations present in our validation sample, which are present in all three subsamples). For all three subsamples, we performed SVD with k=5,000 and calculated distances between all patents and publications within the samples[31].

This results in the same kind of information as described earlier when discussing aggregated results, except for a smaller sample. This means that we can again plot distributions and compare similarity scores of patents and their closest publications for the biotechnology patents and control group patents.

Figure 4 shows the distribution of similarity scores for the similarity measure using TF-IDF weighting and SVD of rank 5,000. The distributions are more shifted to the left compared to the distributions for TF-IDF with lower rank SVD, and there is also a clear distinction between biotechnology patents and control set patents. In short, these distributions are more in line with the expectations, and with the results of TF-IDF weighting without SVD or the similarity measures based on the number of common terms.

[31] Going beyond 5,000 retained dimensions/concept when deriving an SVD solution from our 5% sample takes an extremely amount of computing time and was not feasible for more than one variant.

Figure 4 : Distribution of similarity scores of patents to closest publication according to TF-IDF SVD 5000 (based on 5% sample) (markers=median values)



For the raw document-by-term matrix, and the one after IDF weighting, we find the same kind of results when applying SVD with rank 5,000. In short, we find realistic distributions when SVD with a high number of retained dimensions/concepts is used. Remind that the selection of patents and publication is different for the three samples, so we observe the same kind of improvement for 3 independent subsets. As all validated patent-publication combinations are present in all three subsets, we can again check the congruence between the obtained similarity scores according to those three measures and the expert scores, as we did in earlier.

Table 8 is an extension of Table 3 for the three weighing variants for which we derived a 5% sample and calculated a rank-5,000 SVD solution, and contains again the results of the ANOVA-type of analysis to check consistency between the expert scores and calculated similarity scores. We see that higher number of retained dimensions/concepts have a significant positive effect; for IDF weighting and TF-IDF weighting, obtained results even approach the variants without SVD.

Table 8 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures, including high rank-*k* SVD based on 5% sample (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)

Measure		R^2	Measure		R^2
RAW	**No SVD**	***0.61***	**TF-IDF**	**No SVD**	***0.71***
	SVD 5000 (5% sample)	***0.56***		**SVD 5000** (5% sample)	***0.68***
	SVD 1000	0.34		**SVD 1000**	0.45
	SVD 500	0.31		**SVD 500**	0.34
	SVD 300	0.30		**SVD 300**	0.26
	SVD 200	0.31		**SVD 200**	0.21
	SVD 100	0.30		**SVD 100**	0.17
	SVD 25	0.22		**SVD 25**	0.14
	SVD 5	0.11		**SVD 5**	0.11
BIN	*NA*		**IDF**	**No SVD**	***0.80***
				SVD 5000 (5% sample)	***0.79***
				SVD 1000	***0.63***
				SVD 500	***0.57***
				SVD 300	***0.54***
				SVD 200	***0.51***
				SVD 100	0.49
				SVD 25	0.46
				SVD 5	0.21
Common terms (weighted by min number of terms)					***0.82***
Common terms (weighted by max number of terms)					***0.68***
Common terms (weighted by avg number of terms)					***0.75***

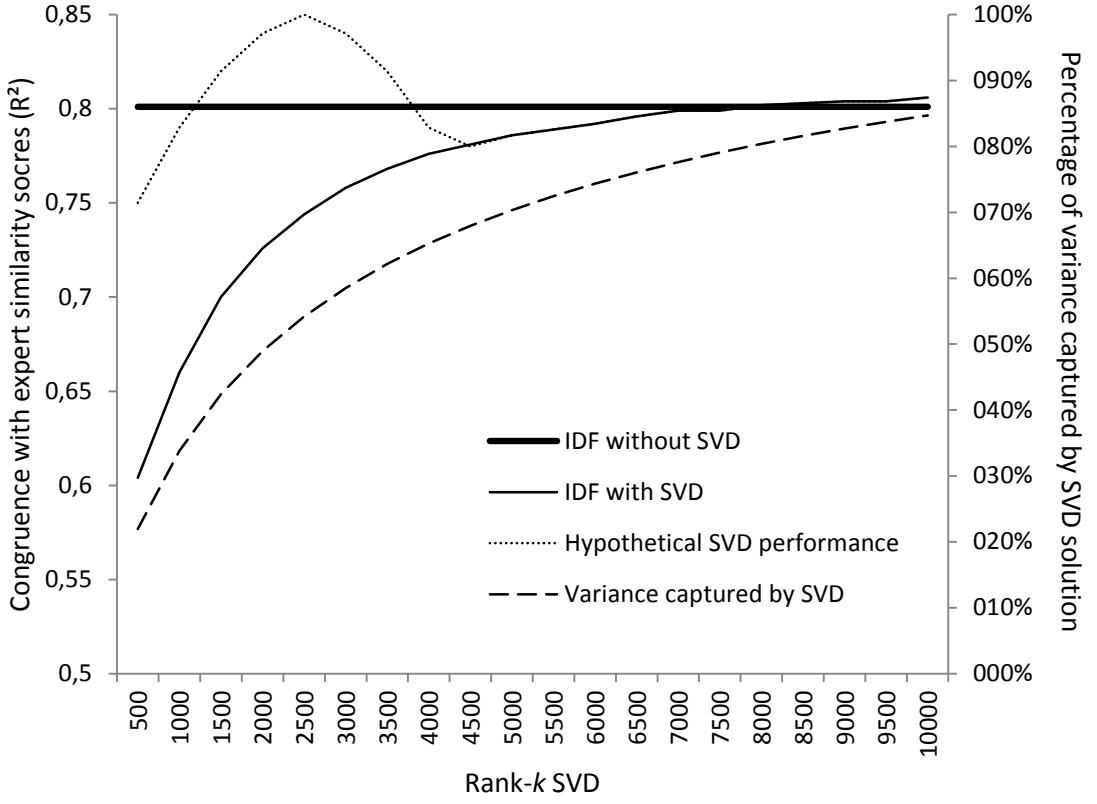
Mean R^2 values in bold denote values higher than 0.5.

To be sure obtained R^2 values are not the result of an accidental good fit of the particular SVD solutions, we created a second 5% subsample based on IDF weighting, derived another rank-5000 SVD solution, calculated similarities of validated patent-publication combinations again and checked congruence with the expert scores, and obtained an R^2 value of 0.793, remarkably close to the R^2 value of 0.786 of the first 5% subset based on IDF weighting. This suggest that SVD solutions are rather stable for a given rank-*k* solutions for subsets within a large document collection.

Table 8 also suggest a positive relationship between the number of retained dimensions and the congruence with expert scores. Figure 5 contains more ANOVA results and lists the obtained R^2 value for the measure based on IDF weighting in combination with high

level rank-*k* SVD in the range 500 to 10,000 in steps of 500 for the 5% subset, together with the captured variance[32].

Figure 5 : Congruence between (conservative) 5-level scale expert similarity assessment and calculated similarity measures based on IDF weighting for high rank-*k* SVD based on 5% sample (R^2 values of GLM regression based on conservative expert scores of 250 validation cases)



The solid horizontal thick line represents the R^2 value obtained when all dimensions are taken into account - cosine on the full vector space defined by the 5% sample after IDF weighting (left axis, 0.801). The thin solid line represents the R^2 value obtained by a given rank-*k* SVD solution based on IDF weighting (left axis, R^2 of 0.60 for 500 dimensions to 0.806 for 10,000 dimensions). The more dimension/concepts are retained, the more the R^2 values of the SVD-based measure approach the one for the full vector space. The dashed line represents the percentage of variance captured by a given rank-*k,* i.e. the variance present in the approximate document-by-term (document-by-concept) matrix after SVD compared to the total variance in the original

[32] We were only able to go to *k*=10,000 for one variant because of the extreme amount of computing time required.

document-by-term matrix of the 5% sample (right axis, 22% for 500 dimension to 85% for 10,000 dimensions).

This table again confirms the positive relationship between the number of retained dimensions/concepts and the congruence with expert similarity scores, approaching the R^2 of 0.801 of a plain cosines on the full vector space based on IDF weighting of the 5% sample.

Striking is the resemblance between the obtained R^2 value from the original dataset and the 5% sample for the same absolute level of dimensionality reduction. Both for the original large dataset as for the 5% sample, the obtained R^2 value when applying the cosine measure to the full vector space is equal to 0.80. The same for the rank-500 and rank-1,000 SVD solutions: R^2 of 0.60 for rank-500 SVD from the 5% sample compared to 0.57 from the full dataset, and R^2 of 0.66 for rank-1,000 from the 5% sample compared to 0.63 from the full dataset. Again an indication that SVD solutions are rather stable for the same rank-*k* level regardless of the data sample used to derive the SVD solution within a large dataset. Mind that the rank-1,000 SVD solution from the full dataset maps 301,697 (stemmed) terms to 1,000 concepts, a reduction to 0.33% of the original number of dimensions. As the 5% sample subset contained 48,561 (stemmed) terms, a rank-1,000 SVD solution from that sample represents a reduction to 2% of the original dimension (capturing 33% of the original variance). It seems that the absolute number of retained dimensions is more important than the relative number of retained dimensions, i.e. relative to the total number of dimensions in the original vector space.

More important, Figure 5 does shed some light on the most pressing question: does a further increase in the number of retained dimension/concepts allow the SVD-based measure to perform better compared to the cosine measure applied on the full vector space, or are obtained similarity scores – and hence performance – merely converging to the ones obtained by the cosine measure on the full vector space. By definition obtained scores of the SVD-based measure will be identical to cosines scores obtained from the full vector space for levels of rank-*k* solutions approaching the original number of dimensions. The point of LSA/SVD is that there are intermediate levels of dimensionality reduction where the SVD-based measure will perform better compared

to the cosine measure obtained from the full vector space. The dotted line in Figure 5 represents hypothetical R^2 values in function of the number of retained dimensions/concepts after SVD as claimed by the LSA method: for levels of dimensionality reduction that are too low, performance will be inferior compared to the cosine measure applied on the full vector space (horizontal thick solid line) because too much relevant information is not taken into account. But for a given range of *k* – in this hypothetical example between *k*=1,500 and *k*=3,500 – performance will be superior because noise – biasing full cosine calculations – is removed from the data, to slide down again beneath the level of the full cosine, to eventually approaching again the performance of the full cosine for values of *k* approaching the original number of dimensions. In our real sample, we do not observe this behaviour, i.e. we do not observe ranges of *k* where the SVD-based measure, based on IDF weighting, performs clearly better than the cosine measure calculated on the full vector space after IDF weighting. We only see the SVD-based measure approaching the full cosine measure. However, there might be ranges beyond 10,000 dimensions/concepts where the SVD-based measure performs better, but unfortunately we cannot check that because of computational limitations to derive those SVD solutions. One can observe at the very right of the figure, for *k*-values beyond 8,000, that the R^2 values of the SVD-based measure are slightly higher compared to the R^2 value for the cosine measure on the full vector space (0.806 versus 0.801), however we strongly believe this is due to rounding errors. Given the curve of the performance of the SVD-based measure in function of the number of retained dimensions, the scenario of (very) high *k*-values resulting in the SVD-based measure to perform better than the cosine measure on the full vector space seems very unlikely; our solution with rank 10,000 already captures 85% of the original variance, and the R^2 curve of the SVD-based measure is almost flat in this region of *k*-values. As SVD solutions are dependent of singular values in descending order, one can expect that the dashed curve representing the captured variance in function of the number of retained dimensions will continue to increase at very slow rates beyond k=10,000, and that the obtained R^2 values will follow this pattern, making further significant increases in R^2 values for the SVD-based measures unlikely. Anyhow, if LSA/SVD indeed requires such very high levels of *k* to perform, and if it would be

feasible to derive such SVD solutions for very high levels of *k* – e.g. by using a sample of documents to derive the SVD solution and projecting or folding in all other documents into the newly created truncated vector space – the method would still be virtually impossible to apply for big datasets because of a lack of storage to retain those big full matrices.

To conclude, we doubt whether further increasing the number of retained dimensions/concepts will result in similarity measures that perform better than a cosine measure derived from the full vector space, let alone the practical feasibility of such a solution. It seems that the dimensionality reduction imposed by SVD is not only cutting off noise, but also relevant information, resulting in the observed pattern of the SVD-based measure approaching but never beating the cosine measures based on the full vector space. This brings us back to the question why this SVD approach would work for some datasets but not for ours.

10 Conclusions, discussion, limitations, and directions for further research

In this study we thoroughly assessed Latent Semantic Analysis (LSA) as a text mining technique to match patent and publication documents based on their contents. The goal was to find patent and publication documents that are related by the topics they address, the methods they use, the results they obtain and the inventions or discoveries they address. This would bypass limitation of current approaches like IPC-codes, non-patent references, and patent inventor and patentee name matching, and allow to compile large scale datasets for a broad range of applications in innovation studies.

As off-the-shelve text mining solutions are not readily available and experience with patent data is limited, we have set up a large comparison exercise based on the LSA method combining four weighting methods and ten levels of dimensionality reduction, and added three measures based on the number of common terms. Our findings reveal that different options and methods available coincide with considerable differences in terms of accuracy. While several combinations allow us to arrive at acceptable

solutions, certain combinations display low levels of accuracy and even result in misleading similarity measures.

Similarity value distributions obtained after application on a large dataset revealed unexpected patterns for LSA-based measures with unrealistic high average similarities and non-biotechnology control set patents being – on average – not less similar to biotechnology publications than biotechnology patents. These results suggest that LSA-based measures tend to overestimate similarity and not grasp the real topic similarity of patent and publication documents.

Expert validation of 250 cases confirmed the poor performance of LSA based measures. SVD dimensionality reduction results in less congruence with the expert assessment of similarity compared to cosine measures applied on the full vector space, and the less dimensions retained, the less congruence. The term weighting method used also effects the performance; binary and IDF weighting yielded better results compared to TF-IDF weighting and no weighting at all, a remarkable observation as TF-IDF in combination with SVD retaining 300-500 dimensions is a commonly used method. We observed that a cosine metric applied on the fill vector space after binary or IDF weighting yields the best results. However, measures based on the number of common terms between documents perform slightly better, in line with Occam's razor principle. The claim that LSA can outperform such simple measures based on common terms or co-occurrence because of a better understanding of the meaning of language of this former method is not backed up by our data.

The weighting method has a significant impact on the performance of the method and it seems that methods taking into account term frequencies perform worse, partly because of stemming and parsing issues, partly because common natural language words tend to get too much weight in the similarity derivation. Better stemming and parsing will probably improve performance.

We propose a combination of measures that allow a more robust identification of similar patent and publication documents: 'common terms MIN', the measure based on the number of common terms weighted for the minimum of the number of terms of the

patent and the publication document, as a primary criterion to identify similar documents, combined with 'common terms MAX', the measure based on the number of common terms weighted for the maximum number of terms of the patent and the publication document, as a secondary criterion to eliminate doubt cases due to combinations of short and long documents. Especially when precision is important, those measures deliver good results. When recall gets important, things get more complicated because there are no threshold values that allow a clear cut distinction between the two groups. The typical trade-off between precision and recall remains a though one, especially as final results are very sensitive to threshold values: small changes in the threshold values for both the primary as secondary criterion result in big differences in the number of matches in the total population. This is particularly problematic for the secondary criterion 'common terms MAX', needed to clear out doubt cases: the vast majority of potential matches based on the primary criterion 'common terms MIN' score very low on 'common terms MAX', so small changes in the range of 'common terms MAX' to discard doubt cases (0.20-0.35) have a huge impact. It seems that our method suffers from too many documents with short abstracts that are very difficult to judge, even for human experts. A potential remedy is to extend document sizes by including patent claims or full documents contents, and not only title and abstract, into the analysis, or use extended abstracts as the ones supplied by the *Derwent World Patent Index*.

When it comes to the identification of patent-publication pairs, i.e. scientific publications from which the contents is covered by patent protection, quality of the results can greatly benefit from an additional third criterion based on the presence or absence of a shared inventor/author. Although inventor-author name matching is not straightforward for larger datasets because of homonymy problems, spelling errors and variation, and use of middle names and initials, the combination of a content based measure like our 'common terms MIN' and 'common terms MAX' and the presence of a shared inventor/author might be the way to go, because the biggest challenge in inventor-author name matching – the homonymy issue – is largely controlled for when combined with a content bases measure.

It is clear that the outlined automated method still have limitations and does not work well in all circumstances. The typical trade-off between precision and recall remains a though one, especially as final results are very sensitive to threshold values: small changes in the threshold values for both the primary as secondary criterion result in big differences in the number of matches in the total population. This is particularly problematic for the secondary criterion 'common terms MAX', needed to clear out doubt cases: the vast majority of potential matches based on the primary criterion 'common terms MIN' score very low on 'common terms MAX', so small changes in the range of 'common terms MAX' to discard doubt cases (0.20-0.35) have a huge impact. It seems that our method suffers from too many documents with short abstracts that are very difficult to judge. On the other hand, this seems not to be due to the shortcomings of an automated method, but due to the characteristics of the dataset. There are simply too many documents with short abstracts that seem to have some relatedness with documents with large abstracts, and even for human experts these combinations are very difficult to assess. We might have to conclude that for these document combinations accurate assessment of similarity is simply impossible. A potential remedy is to extend document sizes by including patent claims or full document contents - and not only title and abstract - into the analysis, or use extended abstracts as supplied e.g. by the *Derwent World Patent Index*.

Regardless of the problems with document combinations that differ largely in size, the number of patent-publication pairs revealed seems low compared to the total population of patents and publications involved. Although precision is high, recall is a problem. The measurement of the thru recall rate in the overall population is difficult. Our validation set does not contain that many missed matches, and hence recall rates seems high, but the global number of patent-publication pairs seems so low compared to the total population of patents and publications that there is no doubt a substantial amount of patent-publication pairs is missed. Ways to improve recall are not clear because we only do have a very limited amount of false negatives in our validation sample. Cross-validation with other datasets with patent-publication pairs would be very valuable to get more insights why patent-publication pairs are missed.

Improving precision and recall levels might be feasible by further broadening the set of pre-processing options. For instance, when inspecting several patent-publication pairs, it became apparent that introducing more synonyms or collocations and phrase detection might further contribute to improving results. More advanced feature selection techniques would also improve the performance of weighting methods that take term frequencies into account, like TF-IDF weighting. Hence, research focusing on the precise impact of additional parameters not included in this design seems relevant. However, practical use might be limited because these additional processing options require considerable manual intervention and this might not be feasible for every single text mining exercise. Unless one would be able to derive synonym lists and collocations that are specific for patents or publications, or that are relevant for a particular science or technology field. One route definitely worth pursuing is the improvement of tokenization and parsing to eliminate errors - especially from chemical formulas which are rather common in our technical dataset - which are getting reinforced because of term frequencies. We believe this will improve the performance of TF-IDF weighting. Related to this issue are stemming errors, also reinforced by term frequencies. One might consider not to use stemming and try to solve synonymy problems using more advanced feature selection techniques.

A remarking observation is the poor performance of SVD-based measures. It is not clear why the specific context of our data does not allow the LSA-method to achieve its full potential. Disentangling the bad performance of SVD in general is very difficult. Looking at individual cases is not very informative as it is virtually impossible to trace back term vectors after SVD to the original terms and contents. The document-by-concept matrix compiled by the SVD solution contains the scores of all documents on newly formed latent concepts, and every latent concept consists of a linear combination of all original terms, i.e. a linear combination with 301,697 components. Although multiple reasons of the limited performance of the LSA application on our dataset can be put forward, hard evidence is limited. It is unlikely that our dataset size is not large enough, nor that we did not retain enough dimensions/concepts. There are indications that the document size and document size differences are negatively influencing the SVD-based measures, and we are almost certain that some of the observed SVD-issues are related to

weighting issues, due to tokenization, parsing and stemming errors reinforced by the use of term frequencies, which we believe is a dataset specific problem caused by chemical formula which are rather common in our technical dataset. But the SVD-issues might also be due to the particular language use in our patent and publication dataset. Again, using the full text of patent and publication documents, or extended abstracts as supplied by the *Derwent World Patent Index*, might resolve this, although we lack hard evidence that larger or better abstracts would resolve the issues.

What is clear is that, for our dataset, the dimensionality reduction imposed by LSA/SVD is cutting off valuable information instead of noise. We observe a gradually increasing performance for increasing number of retained dimensions, but we do not observe a range of dimensions for which the performance is better than that of a cosine measure applied on the full vector space; the performance of LSA/SVD is just approaching the performance of a cosine measure on the full vector space for higher numbers of retained dimensions, in contradiction to the claims of LSA that dimensionality reduction would improve results (understanding the 'latent' structure).

Another remarkable observation is that patents of our materials control set are – on average – more related to biotechnology publications than are biotechnology patents. When SVD-based measures are used. This information can act as a source of inspiration to reveal the shortcomings of SVD on our data. However, looking into many individual cases did not reveal significant information to explain the higher obtained similarity scores nor the poor performance of SVD.

A final reason why LSA might fall short is the limitation to Euclidean geometry as imposed by the assumption of LSA that documents are represented as vectors in a vector space. In an Euclidean space, similarity should be symmetric and not violate triangle inequality - $d(x,z) \leq d(x,y) + d(y,z)$ - placing strong constraints on the location of point in a space given a set of distances (Griffiths, Steyvers & Tenenbaum, 2007). Other text mining techniques not relying on spatial representations, like generative topic models as Probabilistic Latent Semantic Modelling (Hofmann, 1999) and Latent Dirichlet Allocation (Blei, Ng & Jordan, 2003), might be more appropriate to deal with this aspect of the curse of dimensionality.

To conclude, the debate about the value of more complex text mining methods for application on patent and scientific publication data – complex in the sense that they try to deal with the characteristics of text and language – compared to simpler methods based on common terms or co-occurrence does not end here. For our purpose, the identification of patent-publication pairs, a simple measure based on the number of common terms performs best. While claims of LSA are not backed up by our observations, and simpler seems to be better – in line with Occam's razor principle – other text mining techniques are available and it is worthwhile to investigate the application of those techniques on our data, like the generative topic models mentioned before.

11 References

Agrawal, A. & **Henderson**, H. (2002). "Putting patents in context: Exploring knowledge transfer from MIT." *Management Science*, 48 (1) : 44-60.

Argyres, N. S. & **Liebeskind**, J. P. (1998). "Privatizing the intellectual commons: universities and the commercialization of biotechnology." *Journal of Economic Behavior and Organization*, 35 (4) : 427-454.

Arora, A., **Fosfuri**, A. & **Gambardella**, A. (2004). *Markets for Technology. The Economics of Innovation and Corporate Strategy.* Cambridge/London: MIT press.

Atherton, P. & **Borko**, H. (1965). "A test of factor-analytically derived automated classification methods." AIP Report AIP-DRP 65-I.

Azoulay, P., **Ding**, W. & **Stuart**, T. (2009). "The impact of academic patenting on the rate, quality and direction of (public) research output." *Journal of industrial economics*, 57 (4) : 637-676.

Baeza-Yates, R. & **Ribeiro-Neto**, B. (1999). *Modern information retrieval.* New York: ACM Press.

Bassecoulard, E. & **Zitt**, M. (2004). "Patents and publications: The lexical connection." In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 665–694. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Baum, J. A. C., **Calabrese**, T. & **Silverman**, B. S. (2000). "Don't go it alone: Alliance network composition and startups' performance in Canadian biotechnology." *Strategic Management Journal*, 21 (3) : 267–294.

Berry, M. W. (Ed.). (2003). *Survey of text mining*. New York: Springer.

Berry, M. W. & **Browne**, M. (1999). *Understanding search engines: Mathematical modelling and text retrieval*. Philadelphia: Society for Industrial and Applied Mathematics.

Berry, M. W., **Drmac**, Z. & **Jessup**, E. (1999). "Matrices, vector spaces, and information retrieval." *SIAM Review*, 41 : 335-362.

Blei, D. M., **Ng**, A. Y. & **Jordan**, M. I. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3 : 993-1022.

Borko, H. & **Bemick**, M. D. (1963). "Automatic document classification." *Journal of the ACM*, 10 : 151–162.

Callaert, J., **Van Looy**, B., **Verbeek**, A., **Debackere**, K. & **Thijs**, B. (2006). "Traces of prior art: A systematic analysis of other references found within the USPTO and EPO patent system." *Scientometrics*, 69 (1) : 3-20.

Callon, M., **Courtial**, J. P., **Turner**, W. A. & **Bauin**, S. (1983). "From translations to problematic networks—an introduction to co-word analysis." *Social Science Information - Sur Les Sciences Sociales*, 22 (2) : 191–235.

Carroll, J. D. & **Arabie**, P. (1980). "Multidimensional scaling." In M. R. **Rosenzweig** & L. W. **Porter** (Eds.), *Annual review of psychology*: 31 : 607-649. Palo Alto, CA: Annual Reviews, Inc.

Courtial, J. P. (1994). "A coword analysis of Scientometrics." *Scientometrics*, 31 (3) : 251–260.

David, P. A. (2000). "The digital technology boomerang: new intellectual property rights threaten global open science." Working Papers 00016, Stanford University, Department of Economics.

Deeds, D. L. & **Hill**, C. W. (1996). "Strategic alliances and the rate of new product development: An empirical study of entrepreneurial biotechnology firms. " *Journal of Business Venturing*, 11 : 41–55.

Dosi, G. (2000). *Innovation, Organization and Economic Dynamics.* Cheltenham: Edward Elgar Publishers.

Etzkowitz, H. & **Leydesdorff**, L. (1997). "Introduction to special issue on science policy dimensions of the Triple Helix of university-industry-government relations." *Science and Public Policy*, 24 (1) : 2-5.

Faems, D., **Van Looy**, B. & **Debackere**, K. (2005). "Interorganizational collaboration and innovation: Toward a portfolio approach." *Journal of Product Innovation Management*, 223 : 238-250.

Fan, W., **Wallace**, L., **Rich**, S. & **Zhang**, Z. (2006). "Tapping the power of text mining." *Communications of the ACM*, 49 (9) : 77-82.

Freeman, C. (1987). *Technology Policy and Economic Performance.* London: Pinter.

Freeman, C. (1994). "The economics of technical change." *Cambridge Journal of Economics*, 18 : 463-514.

Gans, J. S. & **Stern**, S. (2000). "Incumbency and R&D incentives: Licensing the gale of creative destruction." *Journal of Economics and Management Strategy*, 9 (4) : 485-511.

Glenisson, P., **Glänzel**, W., **Janssens**, F. & **De Moor**, B. (2005). "Combining full-text and bibliometric informationin mapping scientific disciplines." *Information Processing & Management*, 41 (6) : 1548–1572.

Griffiths, T. L., **Steyvers**, M. & **Tenenbaum**, J. B. (2007). "Topics in Semantic Representation." *Psychological Review*, 114 (2) : 211-244.

Grossman, G. M. & **Helpman**, E. (1991). *Innovation and Growth in the Global Economy.* Cambridge: The MIT Press.

Grzybek, P. & **Kelih**, E. (2004). “Anton S. Budilovic (1846–1908): A forerunner of quantitative linguistics in Russia?” *Glottometrics*, 7 (9) : 4–97.

Hall, B., **Link**, A. N. & **Scott**, J. T. (2001). “Barriers inhibiting industry from partnering with universities: Evidence from the advanced technology program.” *Journal of Technology Transfer*, 26 (1-2) : 87-98.

Hardin, G. (1968). “Tragedy of commons.” *Science*, 162 (3859) : 1243-1248.

Harman, D. (1986). “An experimental study of the factors important in document ranking.” In F. **Rabbit** (Ed.), *Association for computing machine’s ninth conference on research and development in information retrieval*. New York: Association for Computing Machines.

Harman, D. (1991). “How effective is suffixing?” *Journal of the American Society for Information Science*, 42 : 7–15.

Hearst, M. A. (1999). “Untangling Text Data Mining.” Proceedings of the 37th annual meeting of the Association for Computational Linguistics: 3-10. College Park, Maryland.

Heller, M. A. (1998). “The Tragedy of the Anticommons.” *Harvard Law Review*, 111 : 621-688.

Hellman, T. (2007). “The role of patents for bridging the science to market gap.” *Journal of Economic Behavior and Organization*, 63 (4) : 624-647.

Hofmann, T. (1999). “Probabilistic latent semantic indexing.” *Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval* : 50-57.

Janssens, F., **Leta**, J., **Glänzel**, W. & **De Moor**, B. (2006). “Towards mapping library and information science.” *Information Processing and Management*, 42 (6) : 1614–1642.

Jardin, N. & **van Rijsbergen**, C. J. (1971). “The use of hierarchic clustering in information retrieval.” *Information Storage and Retrieval*, 7 : 217–240.

Jessup, E. & **Martin**, J. (2001). “Taking a new look at the latent semantic analysis approach to information retrieval.” In M. W. **Berry** (Ed.), *Computational information retrieval*: 121-144. Philadelphia: SIAM.

Kitch, E. W. (1977). “The nature and function of the patent system.” *Journal of Law and Economics*, 20 (2) : 265-290.

Krimsky, S. (2004). *Science and the Private Interest: Has the lure of profits corrupted biomedical research?* Rowman-Littlefield Publishing Co.

Krovets, B. (1995). “Word sense disambiguation for large text databases.” Ph. D. Thesis. Department of Computer Science, University of Massachusetts Amherst.

Landauer, T. K., & **Dumais**, S. T. (1997). “A solution to Plato’s problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge.” *Psychological Review*, 104 (2) : 211-240.

Landauer, T. K., **McNamara**, D. S., **Dennis**, S. & **Kintsch**, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Mahwah (NJ): Lawrence Erlbaum Associates.

Lennon, M., **Pierce**, D. S., **Tarry**, B. D. & **Willett**, P. (1981). “An evaluation of some conflation algorithms for information retrieval.” *Journal of Information Science*, 3 : 177–183.

Leopold, E., **May**, M., & **Paaß**, G. (2004). “Data mining and text mining for science & technology research.” In H. F. **Moed**, W. **Glänzel**, & U. **Schmoch** (Eds.), *Handbook of quantitative science*

and technology research. The use of publication and patent statistics in studies of S&T systems: 187–213. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Leydesdorff, L. (2004). "The university-industry knowledge relationship: analyzing patents and the science base of technologies." *Journal of the American Society for Information Science and Technology*, 55 (11) : 991–1001.

Leydesdorff, L. & **Etzkowitz**, H. (1996). "Emergence of a Triple Helix of University-Industry-Government Relations." *Science and Public Policy*, 23 (5) : 279-286.

Leydesdorff, L. & **Etzkowitz**, H. (1998). "Triple Helix of Innovation: Introduction." *Science and Public Policy*, 25 (6) : 358-364.

Lizza, M. & **Sartoretto**, F. (2001). "A comparative analysis of LSI strategies." In M. W. **Berry** (Ed.), *Computational information retrieval*: 121-144. Philadelphia: SIAM.

Lundvall, B. A. (1992). *National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning.* London: Pinter Publishers.

Magerman, T., **Van Looy**, B. & **Song**, X. (2010). "Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications." *Scientometrics*, 82 (2) : 289-306.

Manning, C. D. & **Schütze**, H. (2000). *Foundations of statistical natural language processing*. Cambridge: MIT Press.

Mansfield, E. (1995). "Academic research underlying industrial innovations: sources, characteristics, and financing." *The Review of Economics and Statistics*, 77 (1) : 55-56.

Mansfield, E. & **Lee**, J. Y. (1996). "The modern university: contributor to industrial innovation and recipient of industrial support." *Research Policy*, 25 : 1047-1058.

McMillan, S., **Narin**, F. & **Deeds**, D. (2000). "An analysis of the critical role of public science in innovation: The case of biotechnology." *Research Policy*, 29 : 1–8.

Merges, R. P. & **Nelson**, R. R. (1990). "On the complex economics of patent scope." *Columbia law Review*, 90 (4) : 839-916.

Moens, M. F. (2006). *Information extraction: Algorithms and prospects in a retrieval context* (The Information Retrieval Series 21). New York: Springer.

Mowery, D. C. & **Nelson**, R. R. (1999). *Sources of Industrial Leadership*. Cambridge: Cambridge University Press.

Murray, F. (2002). "Innovation as Co-evolution of Scientific and Technological Networks: Exploring Tissue Engineering." *Research Policy*, 31 : 1389–1403.

Murray, F. & **Stern**, S. (2007). "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior and Organization*, 63 : 648-687.

Narin, F. & **Noma**, E. (1985). "Is technology becoming science?" *Scientometrics*, 7 : 369–381.

Nelson, R. R. (1993). *National Innovation Systems: A Comparative Analysis.* New York: Oxford University Press Inc.

Nelson, R. R. (1995). "Recent evolutionary theorizing about economic change." *Journal of Economic Literature,* 33 : 48-90.

Nelson, R. R. & **Rosenberg**, N. (1993). "Technical Innovation and National Systems." In R. R. **Nelson** (Ed.), *National Innovation Systems. A comparative Analysis.* New York: Oxford University Press, Inc.

Noyons, E. C. M., **van Raan**, A. F. J., **Grupp**, H. & **Schmoch**, U. (1994). "Exploring the science and technology interface–inventor author relations in laser medicine." *Research Policy*, 23 (4) : 443–457.

OECD (2005). A framework for biotechnology statistics. Paris: OECD publishing.

OECD (2009). OECD Biotechnology Statistics. Paris: OECD publishing.

Ossorio, P. G. (1966). "Classification space: A multivariate procedure for automatic document indexing and retrieval." *Multivariate Behavior Research*, 1 : 479–524.

Porter, M. F. (1980). "An algorithm for suffix stripping." *Program*, 14 (3) : 130–137.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. (www.snowball.tartarus.org/texts/introduction.html).

Porter, A. L. & **Newman**, N. C. (2004). "Patent profiling for competitive advantage. "In H. F. **Moed**, W. **Glänzel** & U. **Schmoch** (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*: 587–612. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ramadan, N. M., **Halvorson**, H., **Vandelinde**, **A**. & **Levine,** S.R. (1989). "Low brain magnesium in migraine." *Headache*, 29 (7) : 416-419.

Rosenberg, N. (1998). "Chemical engineering as a general purpose technology." In E. **Helpman** (Ed.), *General Purpose Technologies and Economic Growth.* MIT Press.

Rothaermel, F. T. & **Deeds**, D. L. (2004). "Exploration and exploitation alliances in biotechnology: A system of new product development." *Strategic Management Journal*, 25 : 201–221.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw-Hill.

Salton, G. & **McGill**, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw Hill.

Salton, G., **Wong**, A. & **Yang**, C.S. (1975). "A vector space model for information retrieval." *Journal of the American Society for Information Science*, 18 (11) : 613-620.

Salton, G. & **Wu**, H. (1981). "A term weighting model based on utility theory." In R. N. **Oddy**, S. E. **Robertson**, C. J. **van Rijsbergen** & R. W. **Williams** (Eds.), *Information retrieval research*: 9–22. Boston: Butterworths.

Sparck Jones, K. (1971). *Automatic keyword classification for information Retrieval*. London: Buttersworth.

Swanson, D. R. (1986). "Fish Oil, Raynaud's syndrome, and undiscovered public knowledge." *Perspectives in Biology and Medicine,* 30 : 7-18.

Swanson, D. R. (1988). "Migraine and magnesium: Eleven neglected connections." *Perspectives in Biology and Medicine,* 31 : 526-557.

Swanson, D. R. (1990). "Somatomedin C and arginine: Implicit connections between mutually-isolated literatures." *Perspectives in Biology and Medicine,* 33 : 157-186.

Swanson, D. R. & **Smalheiser**, N. R. (1997). "An interactive system for finding complementary literatures: a stimulus to scientific discovery." *Artificial Intelligence*, 91 : 183-203.

van Rijsbergen, C. J., **Robertson**, S. E. & **Porter**, M. F. (1980). *New models in probabilistic information retrieval*. British Library Research and Development Report, No. 5587. London: British Library.

Verbeek, A., **Callaert**, J., **Andries**, P., **Debackere**, K., **Luwel**, M. & **Veugelers**, R. (2002). "Science and Technology Interplay – A Modelling Approach on a Regional Level" Final Report to the EC DG Research, Brussels (also forthcoming in the EC Indicators report 2003).

Vidhya, K. A. & **Aghila**, G. (2010). "Text Mining Process, Techniques and Tools: an Overview." *International Journal of Information Technology and Management*, 2 (2) : 613-622.

Wong, S. K. M. & **Yao**, Y. Y. (1995). "On modeling information retrieval with probabilistic inference." *ACM Transactions on Information Systems*, 13 (1) : 69–99.

Wyllys, R. E. (1975). "Measuring scientific prose with rank-frequency (''Zipf'') curves: A new use for an old phenomenon." *Proceedings of the American Society for Information Science*, 12 : 30–31.

Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Cambridge: Addison-Wesley.

Zucker, L. G. & **Darby**, M. R. (2001). "Capturing technological opportunities via Japan's star scientists: Evidence from Japanese patents and products." *Journal of Technology Transfer*, 26 (1-2) : 37-58.

Appendix 1 : OECD biotechnology IPC codes (OECD, 2005 and 2009).

IPC codes	Title
A01H 1/00	Processes for modifying genotypes
A01H 4/00	Plant reproduction by tissue culture techniques
A61K 38/00	Medicinal preparations containing peptides
A61K 39/00	Medicinal preparations containing antigens or antibodies
A61K 48/00	Medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases; Gene therapy
C02F 3/34	Biological treatment of water, waste water, or sewage: characterised by the micro-organisms used
C07G 11/00	Compounds of unknown constitution: antibiotics
C07G 13/00	Compounds of unknown constitution: vitamins
C07G 15/00	Compounds of unknown constitution: hormones
C07K 4/00	Peptides having up to 20 amino acids in an undefined or only partially defined sequence; Derivatives thereof
C07K 14/00	Peptides having more than 20 amino acids; Gastrins; Somatostatins; Melanotropins; Derivatives thereof
C07K 16/00	Immunoglobulins, *e.g.* monoclonal or polyclonal antibodies
C07K 17/00	Carrier-bound or immobilised peptides; Preparation thereof
C07K 19/00	Hybrid peptides
C12M	Apparatus for enzymology or microbiology
C12N	Micro-organisms or enzymes; compositions thereof
C12P	Fermentation or enzyme-using processes to synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture
C12Q	Measuring or testing processes involving enzymes or micro-organisms; compositions or test papers therefor; processes of preparing such compositions; condition-responsive control in microbiological or enzymological processes
C12S	Processes using enzymes or micro-organisms to liberate, separate or purify a pre-existing compound or composition processes using enzymes or micro-organisms to treat textiles or to clean solid surfaces of materials
G01N 27/327	Investigating or analysing materials by the use of electric, electro-chemical, or magnetic means: biochemical
G01N 33/53*	Investigating or analysing materials by specific methods not covered by the preceding groups: immunoassay;
G01N 33/54*	Investigating or analysing materials by specific methods not covered by the preceding groups: double or second antibody: with steric inhibition or signal modification: with an insoluble carrier for immobilising immunochemicals: the carrier being organic: synthetic resin: as water suspendable particles: with antigen or antibody attached to the carrier via a bridging agent: Carbohydrates: with antigen or antibody entrapped within the carrier
G01N 33/55*	Investigating or analysing materials by specific methods not covered by the preceding groups: the carrier being inorganic: Glass or silica: Metal or metal coated: the carrier being a biological cell or cell fragment: Red blood cell: Fixed or stabilised red blood cell: using kinetic measurement: using diffusion or migration of antigen or antibody: through a gel
G01N 33/57*	Investigating or analysing materials by specific methods not covered by the preceding groups: for venereal disease: for enzymes or isoenzymes: for cancer: for hepatitis: involving monoclonal antibodies: involving limulus lysate
G01N 33/68	Investigating or analysing materials by specific methods not covered by the preceding groups: involving proteins, peptides or amino acids
G01N 33/74	Investigating or analysing materials by specific methods not covered by the preceding groups: involving hormones
G01N 33/76	Investigating or analysing materials by specific methods not covered by the preceding groups: human chorionic gonadotropin
G01N 33/78	Investigating or analysing materials by specific methods not covered by the preceding groups: thyroid gland hormones
G01N 33/88	Investigating or analysing materials by specific methods not covered by the preceding groups: involving prostaglandins
G01N 33/92	Investigating or analysing materials by specific methods not covered by the preceding groups: involving lipids, *e.g.* cholesterol
* Those IPC codes also include subgroups up to one digit (0 or 1 digit). For example, in addition to the code G01N 33/53, the codes G01N 33/531, GO1N 33/532, etc. are included.	

Appendix 2 : Example of a patent-publication combination with high but misleading similarity according to the measure based on TF-IDF and SVD.

Following patent-publication combination is an example of a combination with high similarity scores according to the measures based on TF-IDF and SVD. Similarity scores for these measures range from 0.928 to 0.995 depending of the number of dimensions retained (see last line in the table). Title and abstract of both documents make clear that both documents are only (very) slightly related; both are about milk, but the patent is about an apparatus for milking, while the publication is about a comparison of cow milk and camel milk for characteristics on Lactobacillus acidophilus fermentation.

The measures based on the number of common terms yield low scores (0.10, 0.07 and 0.08 depending whether the minimum number of terms, the maximum number of terms or the average number of terms of both documents is used as weighting factor).

Biotechnology patent title and abstract:

Process and rotary milking parlor for the identification of a milking stall and an animal, in particular a cow, in a rotary milking parlor.

For the determination of the occupancy of a milking stall by an animal, in particular a cow, in a rotary milking parlor with a plurality of milking stalls which are disposed on a rotatable milking platform, a process is proposed in which the identification of the animal only takes place after it has entered the milking stall in which it is supposed to be milked.

Biotechnology publication title and abstract:

Growth-behavior of Lactobacillus-Acidophilus and biochemical characteristics and acceptability of Acidophillus milk made from camel milk.

Acidophilus milk was made from camel milk and compared to that made from cow milk. Although the camel milk supported the growth of L. acidophilus, the quality of acidophilus milk from cow milk was superior. Bovine acidophilus had firm curd while that made from camel milk had flocks with no curd formation. The initial proteolysis of raw camel milk provided ready substrates for L. acidophilus for more protein breakdown in the acidophilus milk made from it.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.511	0.837	0.873	0.905	0.754	0.391	0.368	0.608	0.691	0.673
Binary	0.083	0.057	0.025	0.023	0.056	0.087	-0.030	0.492	0.763	0.750
IDF	0.095	0.168	0.162	0.260	0.375	0.403	0.504	0.532	0.698	0.738
TFIDF	0.364	0.928	0.973	0.986	0.991	0.991	0.995	0.980	0.959	0.960

Appendix 3 : Example of a patent-publication combination of a control set patent and biotechnology publication with high but misleading similarity according to the measure based on the number of common terms weighted by the minimum of the number of terms of both documents ('common terms MIN').

9 common terms out of 11 terms of the patent document and 141 terms of the publication document (after stemming, stop word removal and removal of terms only appearing once in the document set): common terms min = 0.82; common terms max = 0.06.

Control patent title and abstract:

Inbred maize line PHBG4

An inbred maize line, designated PHBG4, the plants and seeds of inbred maize line PHBG4, methods for producing a maize plant produced by crossing the inbred line PHBG4 with itself or with another maize plant, and hybrid maize seeds and plants produced by crossing the inbred line PHBG4 with another maize line or plant.

Biotechnology publication title and abstract:

Major QTLs for disease resistance and other traits identified in recombinant inbred lines from tropical maize hybrids

Major QTLs (quantitative trait loci) with large genetic effects often provide the basis for rapid genetic gains with quantitative traits like disease anti pest tolerance. This study sought to identify major QTLs in maize through the creation and use of recombinant inbred lines (RILs) based uniquely on hybrids of elite tropical and temperate inbreds. Nine single crosses involving ten inbreds served as the source of 1072 RILs created through six cycles of single seed descent in the absence of selection in Hawaii. About 30 sublines of each of the ten parental inbreds were bred to estimate means and variances of quantitative traits under study. These parameters were then used to predict RIL segregations of major QTLs based on normal probability distributions, designated here the RIL-NP method. Segregations were also tested for fit to expected ratios by the use of maximum likelihood estimators. The nine sets of RILs were grown selectively under disease epiphytotics at experimental stations in the United States, Korea, Mexico, Nigeria, and the Philippines. Major QTLs apparently acting monogenically (segregating 1:1 in RILs) were identified to control general resistance to the following diseases: Southern rust: Common rust, Northern le:lf blight, Southern leaf blight, Bacterial leaf blight, Stewart's bacterial wilt, Maize mosaic virus and Maize streak virus. Digenic segregations with additive gene action appeared to characterize QTLs governing resistance to Striga witchweed and to European corn borer. Major QTLs were also observed for polymorphisms in ear height, plant height, maturity, tassel branch number and central tassel-spike length. Examples are cited of molecular mapping based on these RILs. The potential use of major QTLs in marker-assisted selection is discussed in relation to the transfer to temper;lce germplasm of tolerances to disease, insect and stress from the largely untapped tropical germplasm.

Appendix 4 : Example of a patent-publication combination with stemming error with high impact on weighting methods including term frequencies.

Both documents have nothing in common, yet score significantly higher for measures based on weighting methods including term frequencies). Both documents have only two (stemmed) terms in common (after stemming and stop word removal), 'feed' and 'ga'. But the stemmed term 'ga' occurs 9 times in the patent document and 29 times in the publication document, resulting in high weights when the term frequency is included. But the stemmed term 'ga' in the patent document is a stemming error derived from 'gas', while the stemmed term 'ga' in the publication document is an abbreviation of 'gibberellin' and has nothing to do with the stemmed term 'ga' in the patent document. For weighting methods not taking term frequency into account, this stemming error counts as just matching term, but for weighting methods using term frequency, this stemming error is magnified and leads to erroneous results.

Biotechnology patent title and abstract:

Incubator with external gas feed.

An incubator with an external gas feed is disclosed, wherein a gas is supplied to an interior space of the incubator to maintain an interior atmosphere with a constant gas-to-air ratio. The gas is supplied to the interior space through a gas nozzle forming a gas jet. The gas jet draws in the interior atmosphere through an injector effect, thereby thoroughly mixing the gas with the interior atmosphere.

Biotechnology publication title and abstract:

Gibberellin metabolism in suspension-cultured cells of raphanus-sativus.

Gibberellin A(1) (GA(1)), GA(4), GA(9), GA(19) and GA(20), which are known to be native to Japanese radish (Raphanus sativus), were applied as [H-3]GAs and [H-2]GAs to cell suspension cultures of R. sativus. As the metabolites in [H-2]GA-feeds, [H-2]GA(8) from [H-2]GA(1), [H-2]GA(1) and [H-2]GA(2) from [H-2]GA(4), [H-2]GA(1), [H-2]GA(4) and [H-2]GA(20) from [H-2]GA(9), [H-2]GA(20) from [H-2]GA(19), and [H-2]GA(1) and [H-2]GA(20)-15-ene from [H-2]GA(20) were identified by GC-SIM. The distribution of [H-3]GA metabolites after HPLC corresponded closely with that of the [H-2]GA metabolites, except in the case of the [H-2]GA(20)-feeds. Based on the metabolic patterns of applied GAs, it is supposed that 13-hydroxylation from GA(4) is much more dominant than 3 beta-hydroxylation from GA(20) in pathways leading to GA(1) in suspension cultured cells of R. sativus.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.688	0.881	0.960	0.958	0.638	0.294	0.229	0.362	0.361	0.443
Binary	0.072	0.088	0.017	0.052	0.056	0.066	0.066	0.144	0.307	0.525
IDF	0.056	0.128	0.128	0.171	0.222	0.218	0.220	0.270	0.471	0.689
TFIDF	0.594	0.941	0.972	0.984	0.988	0.936	0.847	0.886	0.928	0.961

Appendix 5 : Example of a patent-publication combination with tokenization and parsing issues with high impact on weighting methods including term frequencies.

Both documents are not related and have only two terms in common: 'alpha' and 'beta' (after stemming and stop word removal). Both of these terms occur a lot in both documents as part of chemical formulas, and these high term frequencies result in higher similarity values for weighting methods based on term frequencies. But the larger chemical formulas these terms are part of, are not related. If would probably be better to parse and index those formulas as one piece, but this is not straightforward.

Biotechnology patent title and abstract:

alpha -mannosidase inhibitors

¢4S-(4 alpha ,4a beta ,5 beta ,6 alpha ,7 alpha ,7a alpha)!-Octahydro-1H-1-pyrindine-4,5,6,7-tetrols and ¢4R-(4 alpha ,4a alpha ,5 alpha ,6 beta ,7 beta ,7a beta)!-octahydro-1H-1-pyrindine-4,5,6,7-tetrols are useful as inhibitors of alpha-mannosidase and are useful immunostimulants, chemoprotective and radioprotective agents and antimetastatic agents.

Biotechnology publication title and abstract:

Taxanes from Taxus mairei

Four new taxane diterpenes, 9 alpha-hydroxy-14 beta-(2-methylbutyryl)oxy-2 alpha, 5 alpha, 10 beta-triacetoxytaxa-4(20), 11-diene, 2 alpha, 5 alpha, 9 alpha, 10 beta, 14 beta-pentaacetoxytaxa-4(20),11-diene, 5 alpha-(cinnamoyl)oxy-7 beta-hydroxy-9 alpha, 10 beta-13 alpha-triacetoxytaxa-4(20), 11-diene and 5 alpha-hydroxy-9 alpha, 10 beta, 13 alpha-triacetoxytaxa-4(20), 11-diene, along with 12 known taxa-4(20),11-dienes, have been isolated from twigs of Taxus mairei and their structures determined by spectral methods. Copyright (C) 1996 Elsevier Science Ltd.

Similarity values

Weighting method	Dimensions retained (SVD)									
	ALL	**1000**	**500**	**300**	**200**	**100**	**50**	**25**	**10**	**5**
Raw	0.705	0.928	0.930	0.946	0.952	0.970	0.985	0.993	0.999	0.997
Binary	0.089	0.248	0.259	0.295	0.328	0.349	0.198	0.159	0.420	0.785
IDF	0.011	0.248	0.298	0.353	0.389	0.545	0.507	0.397	0.412	0.517
TFIDF	0.199	0.935	0.970	0.982	0.988	0.994	0.996	0.998	0.999	0.997